Social Networks and Its Uses in Collaborative Strategies

A Thesis
Presented to
The Academic Faculty

By

Stephen D Burks

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in Public Policy

Georgia Institute of Technology
July 12, 2004

Social Networks and Its Uses in Collaborative Strategies

Approved by:

Dr. Gordon Kingsley, Advisor

Dr. Michael Farmer

Dr. Christine Roch

Date Approved: July 12, 2004

Acknowledgements

I would like to thank the school of Public Policy for giving me the opportunity to study in it over the last calendar year “under the radar”. I would also like to thank Michael Farmer, Gordon Kingsley, and Christine Roch for giving advice on how to shape this thesis. Both professors have advised me for the last couple of years, and it feels great to be able to see all the work I have done in such a nice order. I would also like to thank Cliff Lipscomb and Dave Schoeneck for their technical assistance on this paper.

I would like to thank Marilou Mycko for taking a chance on me four years ago in the school of Electrical and Computing Engineering. If it were not for her kindness I doubt I would have ever gotten into grad school at Georgia Tech in the first place.

I would like to thank Pat Rose and Bobby Wilson for helping me find an excellent GRA job in the ITTL lab at GTRI to help supplement my education with an excellent work and research experience.

I would like to thank Tara Kilfoyle for changing my life and being the loveliest partner in crime a man could ever want. I am sure there will be many more changes to come once the baby comes.

Table of Contents

List of Tables

List of Figures

List of Equations

Summary

Social networks in their most basic form of nodes and un-weighted connections are viewed in two possible lights for policy analysis: as descriptions and as prescriptions. Descriptive statistics tell us which features are important in a social network and which characteristics make a certain arrangement of nodes unique. However, the one thing that descriptive statistics cannot do is give an indication of what "ought to be". This is due to both the unique history of social networks in policy and also because the study of social networks is still in its embryonic stage.

The early utility of social networks was meant to give a graphical representation of how characters in a policy network interrelate (Heclo, 1978). One of the first motivations of social networks individuals in policy was to show graphically that there are usually significantly more actors in a policy process than were thought in the "iron triangle" sense (Heclo, 1977). Other researchers in the last 20 years have used networks to show the organization of characters along a policy task path (Sabatier et all, 1995). By mapping all actors in a single network, it was easier to explain the importance of all supporting individuals in creating a policy.

While un-weighted social networks have value in their explanatory function at a qualitative level, there has not been much work on what it actually means for two characters to be connected to each other in terms of future connectivity. The relationship graphs of policy analysts usually just have an un-weighted connection between two individuals, which could be interpreted to mean several things. Granted, the knowledge that a connection between two individuals existed during a policy task is valuable, but the

importance of the connection is undeterminable from just this data. Obviously there are individuals and connections that are much more important in completing a policy task, and there are connections that are integral to the policy coming into effect.

Furthermore, causality is hard to establish in un-weighted networks. A connection that exists between two individuals in one policy cycle is not guaranteed to exist in another policy cycle. For example, some fringe actors in a policy system may be integral to one process because they have a special set of skills, but in another process they will not be used. While a policy network with un-weighted nodes presents a snapshot of a group of nodes and their related connections, it is impossible to determine from this map how future evolutions will look and behave. Those fringe individuals may be used one time, but in a network with un-weighted nodes they will look just as important as any other connection in the system.

What can be done to rectify this problem is to place a weight on each of the connections and nodes to determine a probability that a node or connection will be made in future evolutions. Graphically, this would look quite similar to the classical network look, except that for every connection there will be a weighting attached. It would also give a better idea as to how future evolutions of the network will behave than classical models. Past work in econometrics focused on the principles of linearity and continuity in forecasting future outcomes based on past performances. In the model of this thesis, a discrete nodal system relies on randomization of variables over one connection, summed over all connections, to predict future outcomes. Trends may hold, but there is no guarantee that even a connection with a high probability will hold when put through the system.

This evolution is dependant not only on the current arrangement and past growth, but also on future potential for growth and sustainability. For example, in an academic network there may be a node representing a research group, and this node may have been producing many research publications over the last years due to a government grant. However, if the grant were to end, there is a chance that the research productivity would diminish in future years unless another funding source were found. If the node were examined from only past and present output, it would appear that this group were strong, but it is only when the node is examined with respect to future output considerations that its total true potential could be determined.

This does not mean that future potential is the most important aspect to consider in weighing social network nodes and connections. In fact, there is a substantial body of research that focuses on how past research output can affect future output of individuals, countries (J. Furman, Michael Porter and S. Stern, 2002, B. Lundvall, B. Johnson, E. Andersen, and B. Dalum, 2002), and all levels in between (R. Nelson, 2003, Rycroft, R. and D. Kash, 1992). Still, since there is no model that accurately predicts future potential for academic researchers, the best that can be done is to classify potential outcomes into different probability regimes based on the available data. (In other words, heuristics are the extant to which one can use this data.) So there are many variables to consider in mapping out the research potential for a researcher or group of researchers.

Because of the extra effort that is needed to determine a node or connection's true weight, there is an abundance of potential work for both data collectors and data analyzers. Policy has seen a dramatic shift in recent years from prescribing a "one size fits all" best practice policies to more malleable policies based on local conditions. This

is motivated by the work of researchers who saw prudence in abandoning the "command and control" policy dictation (Hjern, 1982; Lipsky, 1980; Maynard-Moody, Musheno, and Palumbo, 1990). Likewise, in a graph of nodes and connections, a similar "bottom up" style of modeling should be fashioned to the local conditions to a graph of nodes. Even though a node in a social network graph looks like all other nodes, it does in fact represent an individual or group with individual characteristics and should be modeled appropriately. Data collectors can help to obtain more accurate weights on a system of nodes and connections through extensive figures, and data analyzers can derive more accurate policy implications and results based on stronger information.

The ironic thing is that individually, the extent to which both groups of research can go is to the stage of descriptive statistics, but with combined efforts, a true policy prescription is within reach. Since much of the work in social networks was originally motivated by digital divide problems, which tend to focus on connecting as many people as possible to a network, it makes sense that a true solution to problems using social networks can only be found when an eclectic group of researchers come together.

In this paper, there are three policy scenarios that are explored and discussed. The first scenario comes from a dataset (Castillo, 2000 and 2002) where little information is known about individual nodes and connection weights are placed based on the economic theory of increasing or constant returns. The second dataset was derived by taking a group of academic researchers (without any knowledge beyond co authorship alliances) working on a joint venture and exploring what combined research ventures would be most beneficial for future research outputs. More information concerning individual nodes and connections is given in this dataset, but the weights on connections are still

developed according to rules of economic theory. The final set of data is developed by viewing the same co-authorship alliances as in the second scenario, but instead the data is examined more thoroughly and more accurate maps of author's connection weights are generated.

It is not that there is a sense of laziness in the first two sections of this thesis, because the state of the art in social networks still lies in descriptive statistics and ways of making their data approachable to more people. Work is done on the datasets to simulate how policies could be injected on them and how they would react, which has not been published in any records to date in policy literature. Still, there is much to be said in treating nodes in generalizations based on only their connections, and future work in the area of social networks for policy analysis should consider more "local conditions".

Chapter 1: Social Networks

Terminology and Preliminaries

In this thesis, there are many terms that are not used by any social networks or network policy organization and may be confusing to people who are either unfamiliar with the jargon or people who approach social networks or network policy from another background. First off, the term *graph* will mean any network of nodes in an arranged manner. In this thesis, *graphs* will usually refer to connected *nodes* that represent individuals or firms that share a common interest or feature, such as a common funding source, a citation in an article, or a common co-authorship. In figure 1, there is an example of a social network representing high school dating. Pink nodes represent females and blue nodes represent males, and connections represent a relationship between those individuals. Social networks can also be used to represent sexual contacts (Newman, 2003), food webs (Martinez, 2001), and even Internet sites (Branigan, Burch, Cheswick. 2000).

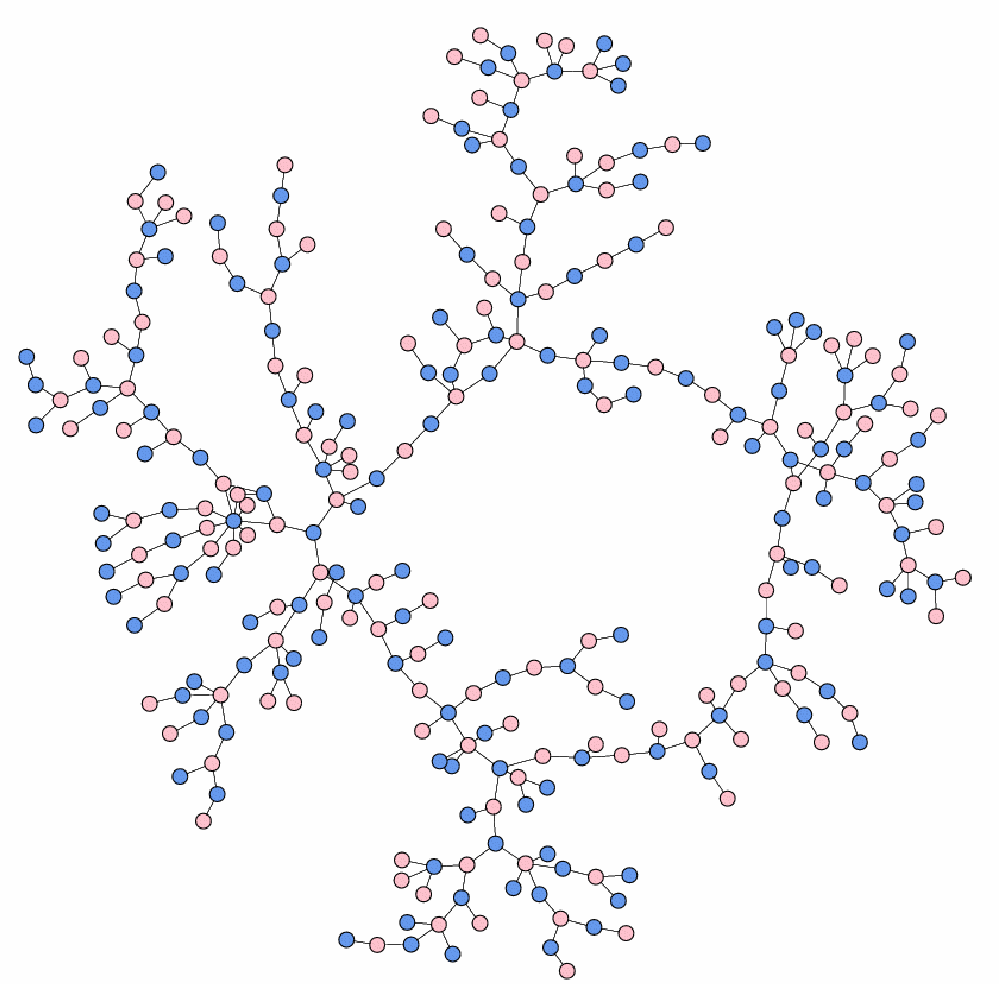

Figure 1: A social network of high school dating. (Moody and White, 2003)

The *matrix representation* of a graph is the representation in which a graph of N nodes is denoted by an N*N matrix. Each row in the matrix represents each node in the system, and in each row values that are nonzero represents a connection to another node. It is of note that nodes have trivial connectivity to themselves, so the diagonal values in the matrix representation will all be one.

It is also noted that connectivity goes both ways, so even though there is only one connection between two nodes it is represented by two non-zero entries in the matrix notation, what some label a bi-graph, but terms differ widely across mathematics,

physics, sociology, computer science and political science. All this means is that, for instance, in the graph below, node 1 is connected to node 2, so the value of the matrix representation for the graph has values of A(1,2)=A(2,1)=1. Therefore, the general notation of a matrix will be an N*N matrix that has diagonal values all equal to one and is also diagonally symmetric so that for all A(i,j)!=0, A(i,j)=A(j,i). Some persons do represent social networks that are unidirectional in representation, but since the policy context and scope of this research is in collaborative strategies and joint ventures, the bi-directional notation is the only one that will be in use.

While they exist, it is very difficult to locate many examples of unidirectional connections in the social network literature as sociological theory strongly emphasizes relational quality. Unfortunately, many of the strategies from computer science in parallel processing or from physics in system dynamics are 'forced,' a technical term that covers several phenomena; but in the graph theory context this often means that associations are unidirectional. One needs to be careful regarding the mathematical assumptions. So it is unfortunate if a social network analyst adopts a tool from physics or from computer science where a 'forced' system with a backward feedback loop is equivocated to a bi-graph without full inspection. In my experience, nearly every social network application that appeal to the theory of neural networks as an analogy for collaborations and associations conflate these two very important technical differences.

Farmer (2004) suggests that policy implementation projects where bureaucrats with resources interact with a social network are conducive to the unidirectional analysis with a feedback loop; yet the presence of hierarchy or central nodal positions itself in a social network (Moody, 2003) is not sufficient to apply the neural network thesis.

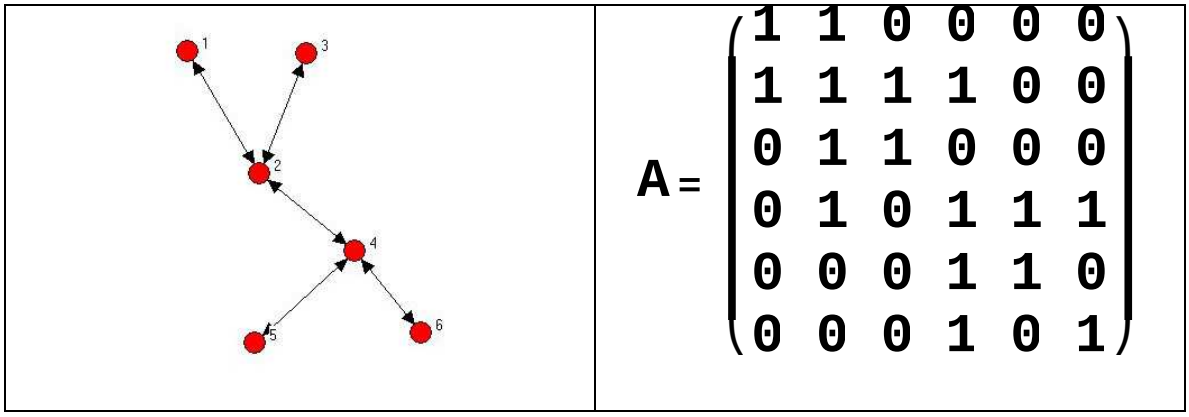

Figure 2: A simple social network and its matrix notation representation.

The matrix notation's utility is not only in its simplicity in determining characteristics of the social network but also in its simplicity in creating and removing nodes and connections. In other words, as research or investment networks grow through success, this format makes it easier to track and compute the likelihood of new entrants and exits.

All that is required to create a connection between two nodes *i* and *j* is to change the values of the matrix from A(i,j)=A(j,i)=0 to A(i,j)=A(j,i)!=0. To add a connection to the matrix, all that is needed is to change the dimensionality of the matrix from N to N+1, change the diagonal element to 1 (A(N+1, N+1)=1), and add connections just is mentioned above. To remove a connection between two nodes *i* and *j*, the values of A(i,j)=A(j,i)!=0 are simply changed to A(i,j)=A(j,i)=0. Nodes will not be removed the system in this research presentation since matrix notation is not conducive to bookkeeping of individual node behaviors.

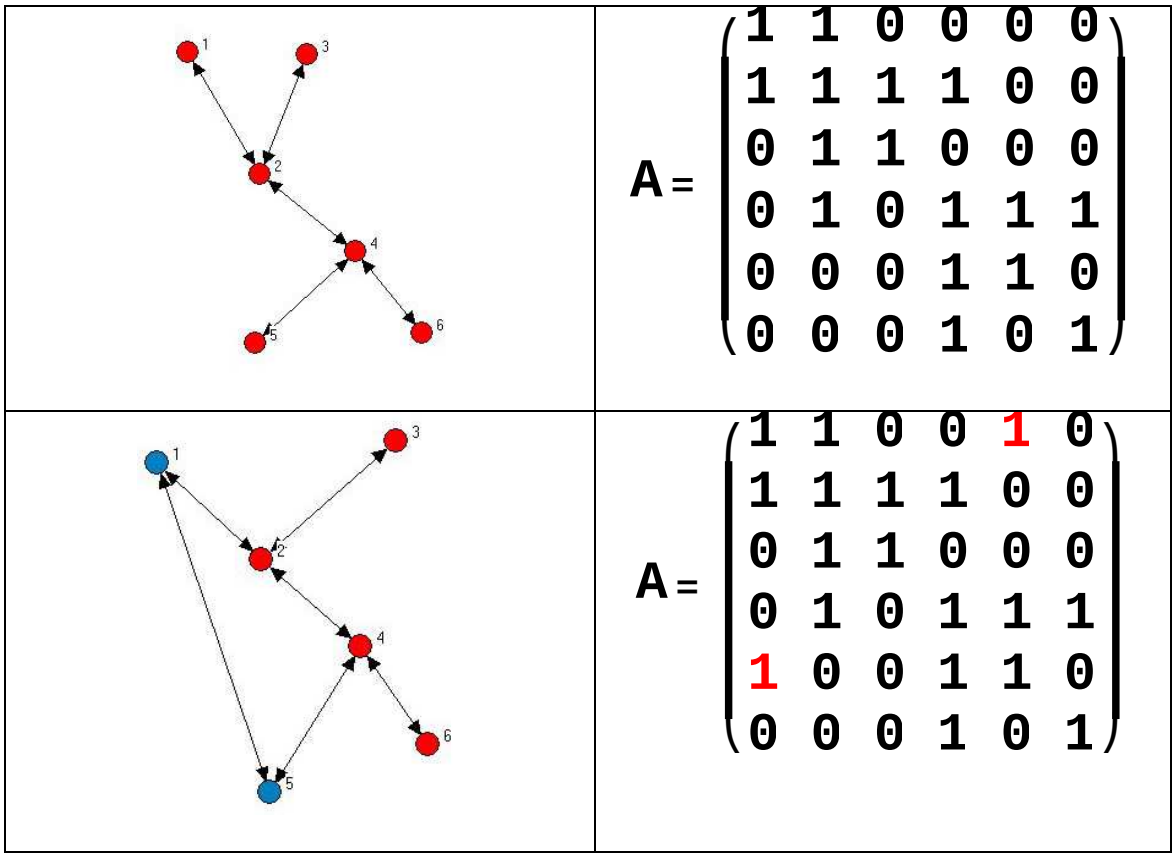

Figure 3: Graphical and matrix representation for the addition of a connection.

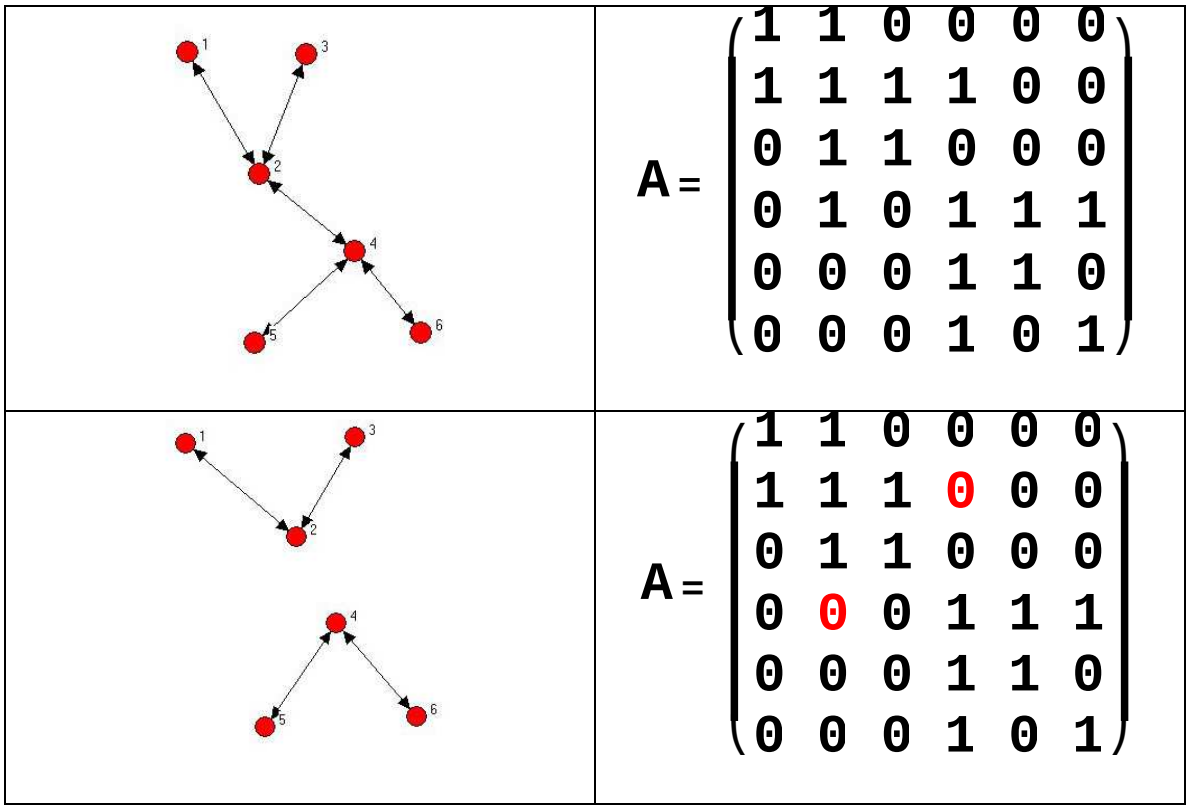

Figure 4: Graphical and matrix representation for the removal of a connection.

The final utility of the matrix notation is that by multiplying the complete matrix by itself, it represents the whole system of nodes connecting (or jumping) to the next reachable node. For instance, if a node is connected to only one node (represented by a row with two non-zero values, one for the connection to itself and one for the connection to the other node) which is connected to three nodes, after multiplying the matrix by itself one time, in the new multiplied matrix the row representing the original node connected to only one other node now will have five non-zero values (the two from the original connection and the three additional ones from the connection from the node). This indicates that the node is reachable to one node other than itself after one jump and is accessible to four other nodes than it self after two jumps. Again the matrix notion is quite flexible to capture important social and policy characteristics that can be attributed

to a network; and this feature of connectivity (what I call accessibility to other nodes for a given nodes) is a central characteristic to policies directed to association probabilities and success. So this definition is important also.

$$Accessibility_i^M = nnz((A^M)_i), \text{ where is } (A^M)_i \text{ is the } i^{th} \text{ row of the } m^{th} \text{ power of A}$$

Equation 1: The Total Number of Accessible Nodes from the i^{th} Node after M Jumps.

This allows us to formalize what I call *reachability; or* the *reachability* of a graph of nodes is the ability of one node to reach another node by means of the connection between nodes. Strictly accessibility is overall reachability after so many jumps (or degrees of separation in some texts). For purposes of discussion here the terms are analogous and the two expressions are used interchangeably in the forthcoming analysis.

Many times in the graphical use of social networks the terms *hub* and *spoke* will be used to describe nodes and connections. The reason for the use of these terms is that the hub of a graph of nodes is considered by many people to be a node of importance because one can travel from this one node to many other nodes in the quickest amount of time (or fewest jumps) through the hub node. Much early social network research focused on finding the hubs of a cluster of nodes. (Burt, 1992)

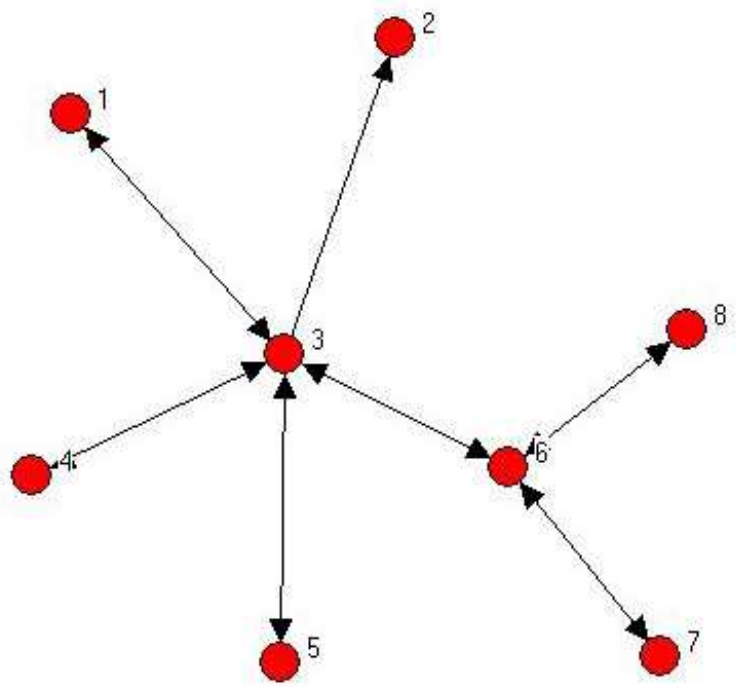

Figure 5: Hub and Spoke Model Example. Node 3 is the major hub of the system while node 6 acts as a secondary hub to the system.

The term *hub*, though turns out to be quite context dependent. For job searches, three jumps is considered quite accessible (Grannovetter, 1974); but in other circumstances, anything other than immediate connectivity is almost useless while in disease spread over a small world graph, 5 or 6 jumps – or degrees of separation – is enough (Watts and Dodd, 2001). If different associations or clusters are more fluid and outward reaching than others, as is the case we suspect with human networks, 'hub' will also depend on *which* nodes or path of connections the person encounters. So the 'hub' of a graph is far from fixed from graph to graph or from task to task, but will change to make the theoretical concept of a 'hub' quite useful (Burt, 1992) but as an immediate analytic instrument, the term is technically quite vague in the current literature.

The phrase *jump* is used loosely in this paper to describe nodal distance. For instance, if two nodes are connected to one another, it is said that the two nodes are

connected by “one jump”. If two nodes are connected through one mediator node, then it is said that the nodes are connected to each other through “two jumps”. A better way to determine the actual distance between individuals is in terms of their probability of connection. For instance, if two individuals have a probability of connection of 0.3, then their distance should be calculated to be 0.3. All distances would be normalized to be between 0 and 1, and the probability of connection for individuals not directly connected but instead connected through intermediaries would be the product of all the probabilities of connection between those individuals.

Of course, the probability will lower geometrically as the number of jumps needed to get from one researcher to another increases. This is one of the motivating factors for the policy injection in the form of the connection between distant or disjoint individuals. Bridge connections between distant or disjoint individuals can make a huge difference in determining the probability of connection between individuals that would otherwise be several jumps from each other, or not connected to each other at all.

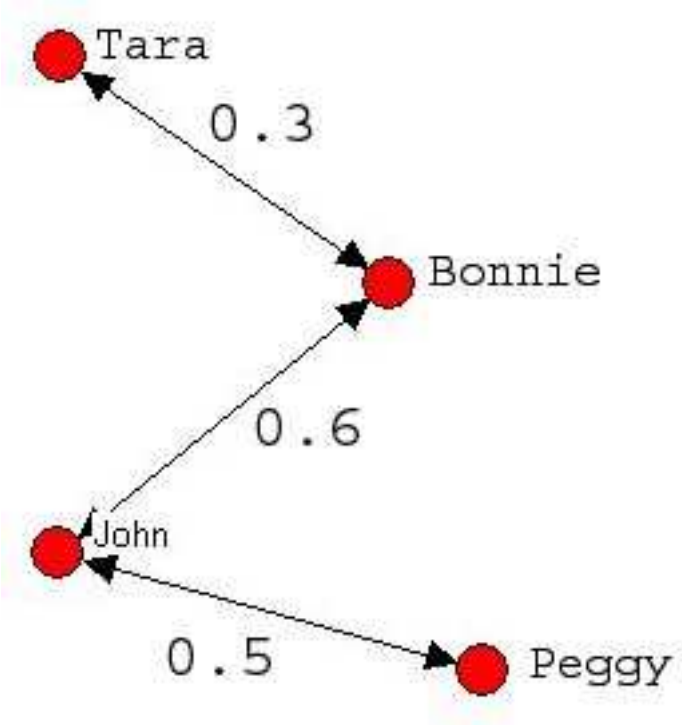

Figure 6: Connections between individuals with probabilities not equal to one. Tara and Bonnie have a connection probability of 0.3. Bonnie and John have a connection probability of 0.6, and John and Peggy have a connection probability of 0.5.

The terms and formulae above are motivated by social network theory itself. Where they depart from terminology or conventions in physics or computer science, it is for this reason alone. Moreover, the review suggests that borrowing from other disciplines for the purpose of social network application can be tricky. The technical terms are defined so that the theoretical distinctions map directly to the techniques and analytic practices deployed in this work. This is a new way to analyze the dynamics of a social network that is at once less technical than other rules from the physical sciences and more accessible and pertinent to the policy questions raised by social network theory.

Chapter 2: Data Set 1 and First Analysis using Social Networks

Introduction

The first analysis will cover data that was taken on companies that were funded by venture capitalists in Silicon Valley. If the same venture capital firm funded two companies a connection was drawn between them. What is not known about this dataset is when these companies were funded or if they had any correspondence with companies with shared venture capital sources. This dataset was originally acquired in a manner to keeping the exposure open on a camera for an extended time (Castillo, 2003). All of the funding was captured over an extensive period of time and the results are made available on a single picture.

This method of data capture has advantages and disadvantages. The biggest advantage is that systematic analysis is possible since all connections are captured. This can be different from some policy networks because this graph will give an indication if two companies have ever had any relationship over a certain period of time. The weakness to this output is that it is static and that there is no determination as to how companies came into the network. Also, as in traditional policy networks, there is no knowledge as to the strength of the connections between companies.

From this static data, assumptions concerning it can lead to predictions on future output through simulations. The model section of this chapter tells exactly how this is accomplished on the Castillo dataset. It is important to note that the model of simulation relies on the static data from the initial picture of the dataset. There is no desertion of

static analysis; rather, the static case acts as a “seed” for future output within a system. This is analogous to the importance of the initial state in highly dynamical systems (Strogatz, 1994)

There are several policy implications that can arise from this data. One is that there is potential for drawing future policy based on a snapshot of past and current activity. Also, if assumptions are held regarding connectivity and its role in innovation, policy will be in the form of simply making or breaking connections between nodes to accomplish a task. The goal of the policy in this section will be that of stimulation, which will be in the form of creating connections between nodes to maximize output, keep as many of the original companies in the system as possible, and bring as many new companies into the system as possible.

The largest policy consideration not yet mentioned is related to communication. The Castillo dataset represents companies that work in the innovation industry, and policy to connect people to work together on a jointly funded project will not only stimulate those two companies that have a high potential for output, but it will increase the diffusion of knowledge throughout the system. For instance, if two companies are connected to each other through several intermediary nodes, there is little to no possibility of knowledge of each other’s work. But if a connection were made between the two there would be direct knowledge transmission between the two. Also, there is a greater chance that the peers of both companies (those that are connected to both) would receive information concerning this otherwise unreachable knowledge base.

Descriptive Statistics of Dataset 1

The first dataset, which was acquired from previous work done by Castillo (Castillo, 2000 and 2003) on joint ventures in Silicon Valley, is unique in that it gives a notion of how an actual social network *looks* and behaves. The original dataset was acquired by putting Castillo's graphical output in the matrix notation. Each node in the Castillo set represents a company, and a connection between two nodes represents a common funding source from the same venture capital firm. In the analysis of this dataset there are no weights because the Castillo article does not give more information on the specifics of each connection between two nodes.

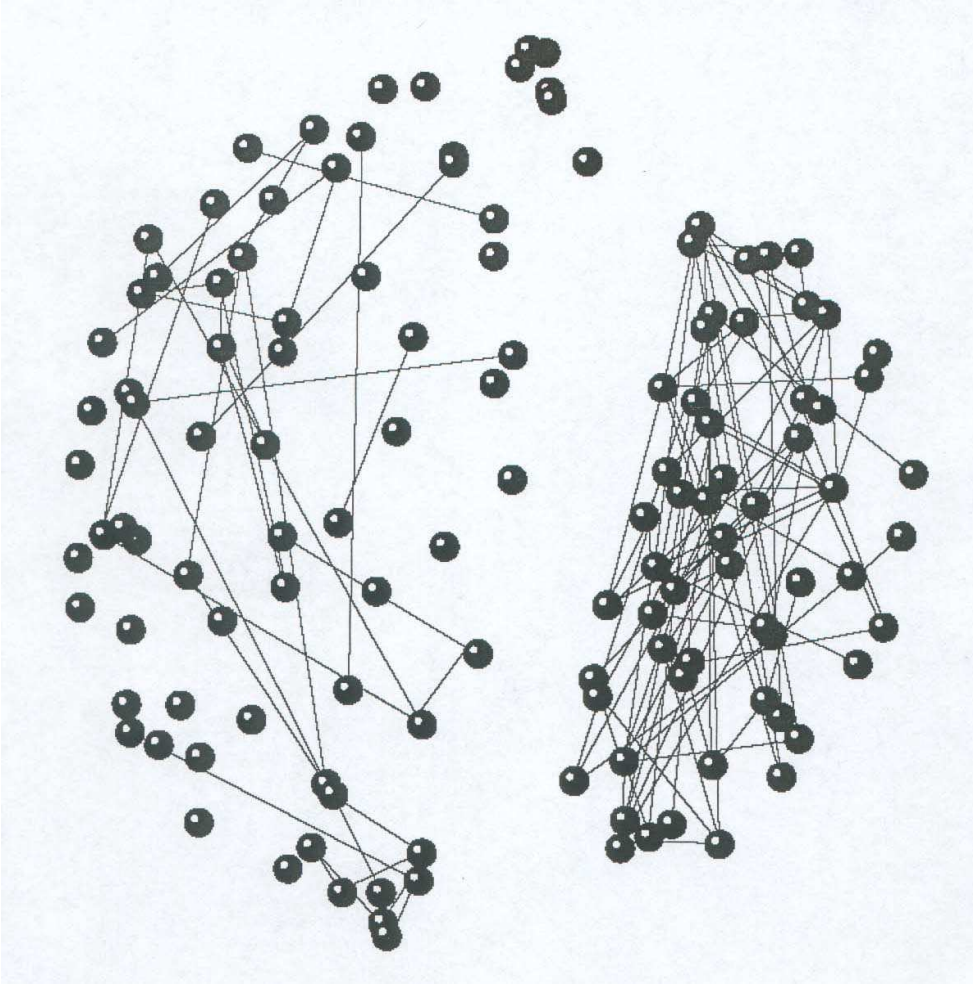

Figure 7: The original dataset as was contained from the Castillo article.

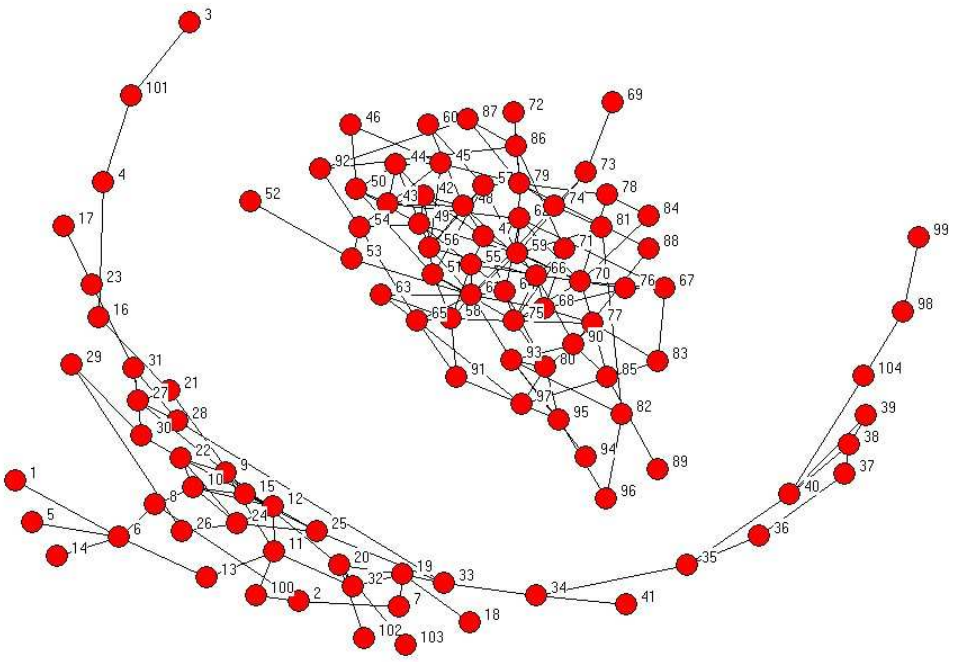

Figure 8: The original dataset after it was converted into matrix notation and graphed with the UCINET program

The Castillo system has a total of 104 nodes, and there are a total of 502 non-zero values for connections in its matrix representation. The number of actual inter-nodal connections in the system was calculated by first subtracting off the trivial connections of a node to itself and then dividing the remaining number by 2, because the connections between nodes go both ways.

$$\text{Connection} = \frac{\text{nnz(A)} - \text{Dim(A)}}{2}$$, where Dim(A) is the dimension of A, or the number of nodes in the system, and nnz(A) represents the total number of non-zero elements in the matrix representation of A of the graph of nodes.

Equation 2: Calculation of the number of connections in the system

It was found that there are 199 inter-nodal connections, which translates to a baseline density of 3.7% if the trivial connections are not kept in the system ($Density_{trivial}$) and 4.6% if the trivial connections are included in the statistics ($Density_{non-trivial}$).

The density gives an indication as to how much connectivity there is between nodes. For example, if all nodes were connected to all other nodes, the density would be 1.0. The equation for calculating the graph density is as follows:

$$Density_{trivial} = \frac{nnz(A)}{Dim(A)^2}$$

Equation 3: Calculation of the Trivial Density of the Graph of the Castillo System

$$Density_{non-trivial} = \frac{\frac{nnz(A) - Dim(A)}{2}}{\frac{Dim(A)^2 - Dim(A)}{2}}$$

Equation 4: Calculation of the Non-Trivial Density of the Graph of the Castillo System

From a casual observation, it is noted that there are two distinct clusters of nodes in the Castillo system. Within each of these clusters of nodes, there is total reachability, meaning that any node can reach any other node in the cluster by means of the connections between the nodes. This is proven analytically by examining the number of non-zero values in the matrix after taking the nodes to their furthest reaching nodes.

$\text{Reach}_i^{Dim(A)/2} = \frac{\sum_i^{M'} nnz((A^{Dim(A)/2})_i)}{M'} = M'$, where *i* is the row index of the system and M' is the size of the cluster of nodes under examination. In this instance, the cluster of nodes is less than the total number of nodes in the system, but that that is not necessary for determining the average reachability of a set of nodes in a system. Since the system is bi-directional, it is only necessary to take the system to half the length, as the nodes will connect in both directions.

Equation 5: Calculation of the reachability of a graph of nodes within a certain cluster.

The average path length of the system of nodes is defined to be the total number of jumps necessary to get from one node to all other nodes, summed over all nodes in the system, and then divided by the total number of nodes. This number is usually represented by the ratio L/C, where *L* represents the length of total reachability within the system or graph for a given number of jumps (also referred to as the path length) and *C* represents the number of nodes in the system or graph. The L/C ratio is difficult to interpret if there is not total accessibility in the system because there will be nodes that are not reachable even after an infinite number of jumps, so there is a choice of either only summing up those nodes that are accessible, or of asserting that if one node is not accessible then the total graph has an undeterminable L/C ratio. In this paper, if there are any nodes that are unconnected then the L/C ratio will be reported as an indeterminate number.

$\text{L/C} = \frac{\sum_i^N TotalJumps_i}{N}$, where N is the total number of nodes in the system and the index *i* is the individual node whose jumps to all other nodes is summed. The code used to find the ratio of L/C is found in the Code section of the Appendix.

Equation 6: Calculation of the ratio of path length to the number of nodes in the system.

A principal components/factor analysis was performed on the Castillo dataset using the UCINET program. The method of analysis is based on searching for similarity in the profiles of distances from each node to others, and then displaying the nodes in a fashion where those nodes with a higher principal component score are graphed to the right of those nodes with a lower principal component score. The results of the principal components/factor analysis were that the nodes on the more densely packed, higher numbered nodes on the right side were more important to the system than the lower numbered nodes on the left side. Those nodes with the highest score were connected to many other nodes in the graph and acted as a hub of spokes to other nodes in the graph.

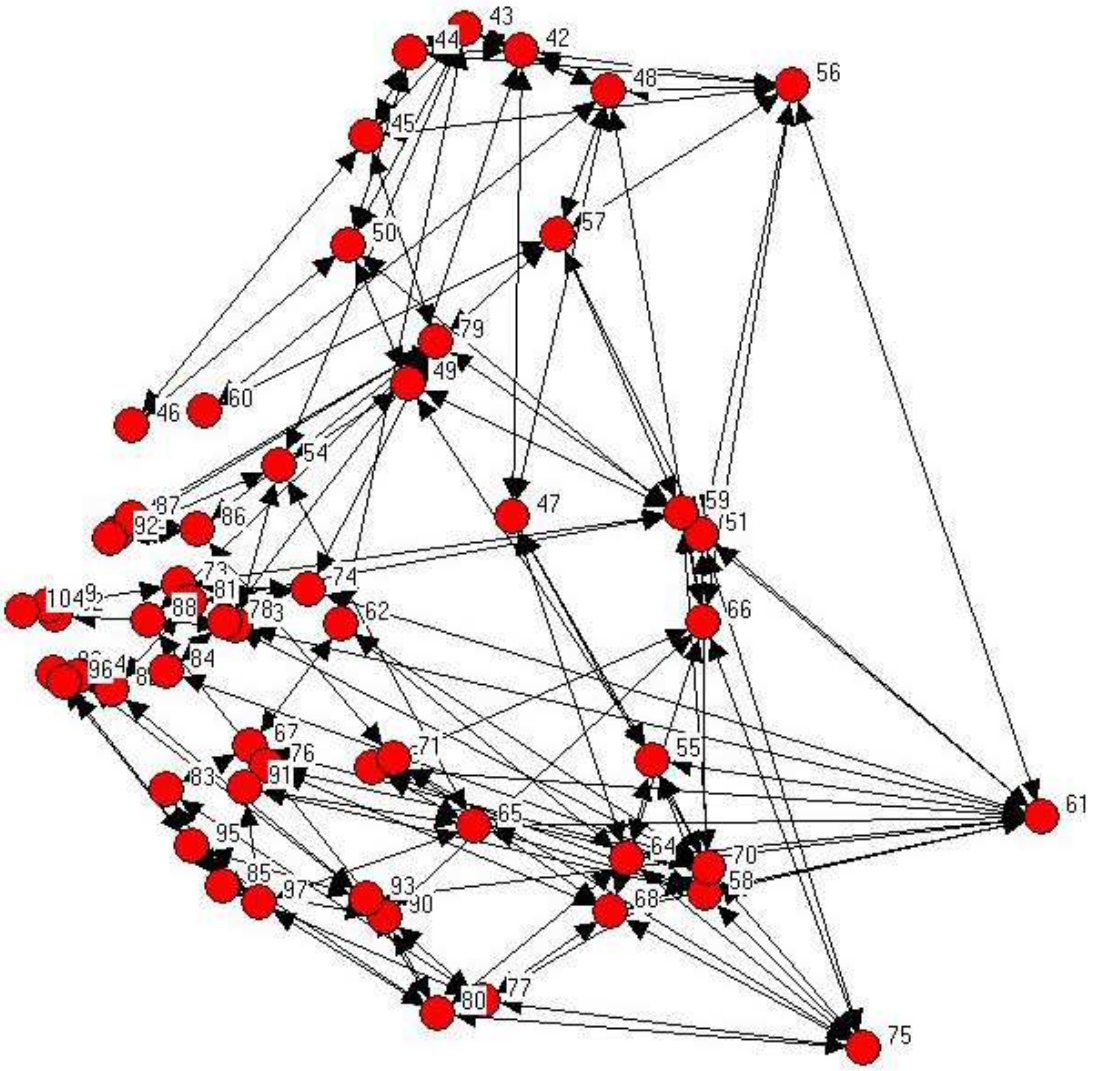

Figure 9: Principal Components layout of the Castillo dataset.

Policy Injection in the Castillo Dataset

The objective of policy injections on the Castillo dataset is to not only keep as many actors (or nodes) in the system after a specified number of policy cycles, but also to maximize the output of the connection collaborations. In all of the forthcoming policy injections, the choice of connection to make on the system is highly dependent on the amount of information regarding the connections and nodes. For instance, system of nodes that has no information concerning those individual node's probability of success may want to rely on a fairness/inclusion policy because that policy makes an attempt to keep as many nodes in the system as possible.

It is noted that success in a policy simulation does not always equate to success in the "real world". The simulations that will be run on these systems are a simplification of the dynamics that occur between actors in a system. A strong success in a policy simulation would not always mean that the prescribed connection would work in an actual policy injection. Likewise, individuals whose collaborative efforts lead to great results might be counterintuitive to what would be chosen in any policy map or simulation.

Because there is no mention of the specifics of each of the nodes of the connections in the Castillo dataset, there is no clear policy prescription that would act as a universal policy of improvement (or "silver bullet") on this system of nodes. The main assumption to each of the policies that will be prescribed onto this dataset is that there is a set level of knowledge concerning the nodes and the connections between them. Some policies assume that there is no knowledge of the nodes or connections, so within a framework of increasing or constant returns the policies will rely more on the philosophy

of how joint ventures work and the topography of the graph. As the amount of information in the graph rises, there will be less dependence on the large-scale features of a graph and more emphasis can be placed on strategic connections between nodes that would be likely to produce the most output.

There are several ways in which one can impose or inject a policy on a system of nodes in a graph, but they all come down to doing one of four things. A policy injection on a network will add nodes, subtract nodes, add connections, or subtract connections. The policy injection that is used in this research is to add connections between nodes. In the case of the Castillo dataset, since there are two unique clusters of data, the logical choice for connectivity will be between nodes on each of the two clusters. Five connections were chosen by means of various policies regimes. As a note, the addition of five new connections into the system raises both the trivial and non-trivial density by only 0.1%, and these additional connections raise the total number of connections in the system by 2.5%. Yet, the reachability in the graph's static state will become 100% for all values in the network.

The main goal of policy in this simulation, as in most other policy scenarios, is to maximize the benefit for the cost. In this case, the benefit will be measured in terms of the number of successful connections made or kept between researchers, and the cost will be the number of funded connections between researchers in different groups. Other measures of benefit will be total reach of connectivity in the system after the simulation has run through one policy cycle. A successful connection represents a successful joint venture in between two researchers. The number of successful connections in this model

will correlate directly with the amount of successful ventures that come from the system's model.

A baseline for the policy choices is used to compare how each of the policies affects the system in comparison to if nothing were done to the system. The baseline acts as both a control and a counterfactual. It is noted that even if one connection were added to bridge the two clusters of nodes, the dynamics of the Castillo system will change dramatically. The baseline for the graph will not have complete accessibility in the static case, so some statistics are not useful in describing the system.

One policy choice that was not implemented was to find those connections that would minimize the L/C ratio for the graph of nodes. To find these connections, there is great computation needed. In the case of the 104-node Castillo system, there were over 5,000 different nodal connection comparisons to make, each one requiring several thousand-matrix multiplications which takes a significant amount of time to compute. It was estimated by Matlab that on a Pentium 3 processor with 512 megs of RAM that it would take over a month of constant computation to find those connections that would minimize the L/C ratio. A policy based on finding the connections that would minimize the L/C ratio would transcend the fairness rule introduced later in this section because it would find the group of connections that would minimize the total path length of all the nodes to all the other nodes.

The L/C policy is probably best placed in the paradigm of the classical problem in policy analysis of efficiency versus equity. The connectivity fairness policy for this section is really not that fair because it only looks for those nodes that have the highest reachability in the fewest number of steps and then connects them to each other across

both sides of the connection divide. The routine needed to generate these connection choices is determined in just a fraction of seconds on a standard computer, but the price of this ease of connection determination is that there is a much higher possibility of nodal alienation in the system. The L/C reduction policy is deontological in that it tries to make the number of jumps of *all* nodes to *all* other nodes as small as possible *on average.* So what sounds fair nests important potential unfairness at the extreme, helping the alienated but picking the low lying fruit rather than making the worst off as well as possible.

The first policy connection choice is based on the idea of connection maximization and is called direct optimization. This rule is based on the idea that the logical choice for connectivity of nodes would be between those nodes that are the hubs of their graphs. (Burt, 1992) This selection process seeks those collaborations between firms that would immediately maximize total output – or highest expected output under the probability regime characterizing the static network map. The process for searching for candidates in the direct optimization is simple in that it involves counting the number of connections from each node to another node, and then connecting those nodes to each other with the highest connectivity.

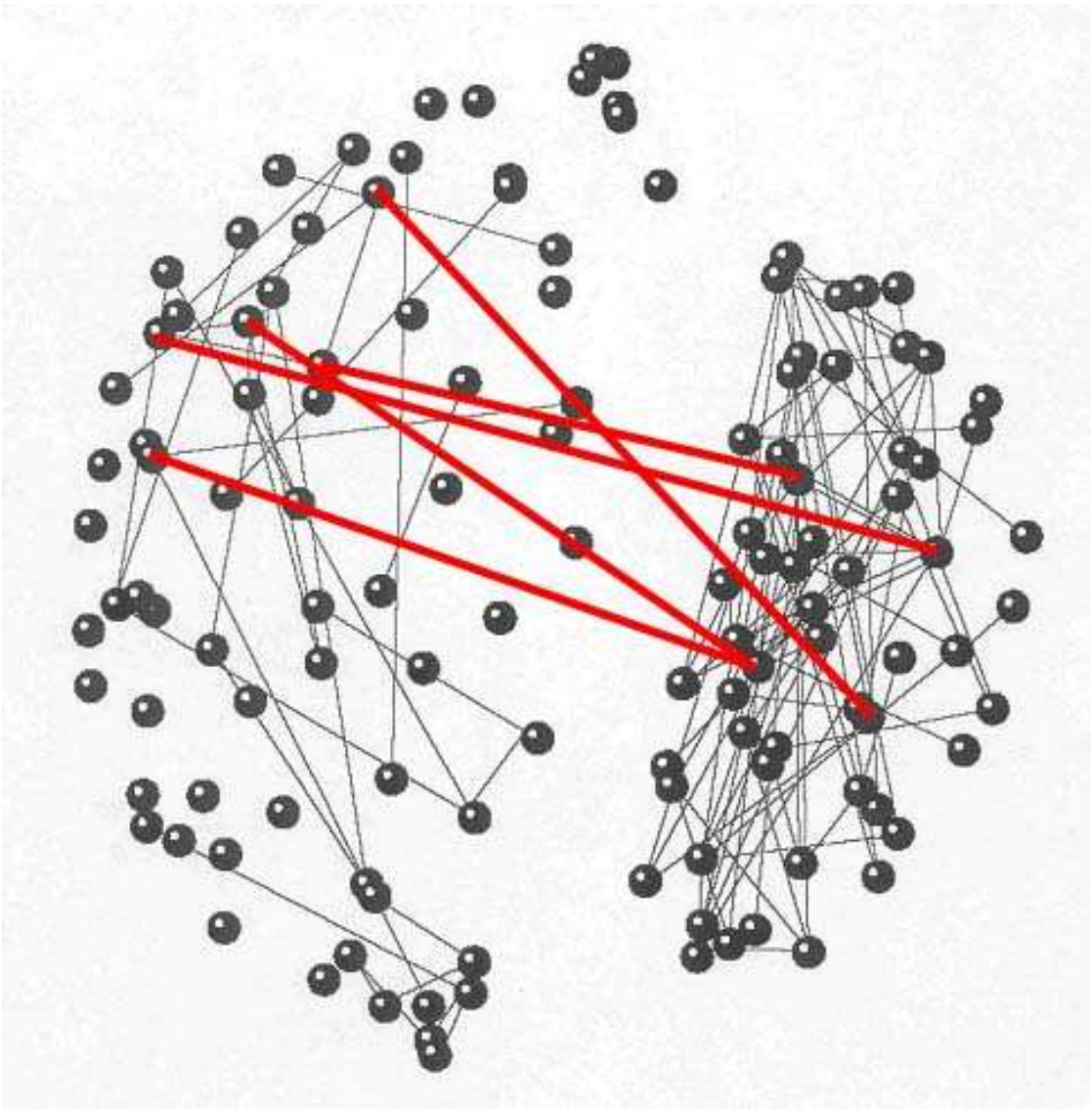

Figure 10: Direct Optimization/Connection Maximization Policy

A second policy selection choice is the so-called Smart Small World rule. This rule connects those nodes that have the highest amount of connectivity two degrees from themselves. This policy comes from information transmission theory in that a person can ask their peers for a piece of information, and their peers will likely know of the work of their other peers but not of their peer's peers. Also, the Smart Small World policy is similar to that of forming a community circle in that there is a familiarity and a higher percentage of tighter connections since there is a commonality of friends. This policy

emphasizes density yet can potentially penalize centrality within the graph. The Smart Small World policy targets nodes that overlap prominent clusters rather than simply link across clusters of well-centered nodes in a cluster that enjoys hierarchic influence. (Moody, 2003). Another way to frame the policy in terms of new growth theory is that the rules aspire to maximize the diversity and the industrial complementarity of monopolistic competition without degenerating into an oligopoly (Fujita, 1993) or coordination between decomposition and centrality (Jackson & Watts, 2002). The method for determining the candidates for the Smart Small World policy is to count all of the connections from each node and also count the connections from those connected nodes. The simple method for doing this is to work in the matrix framework and multiply the matrix by itself, meaning to take the system to its second "jump", and then the values in each column will be the number of nodes that are reachable within two jumps from itself.

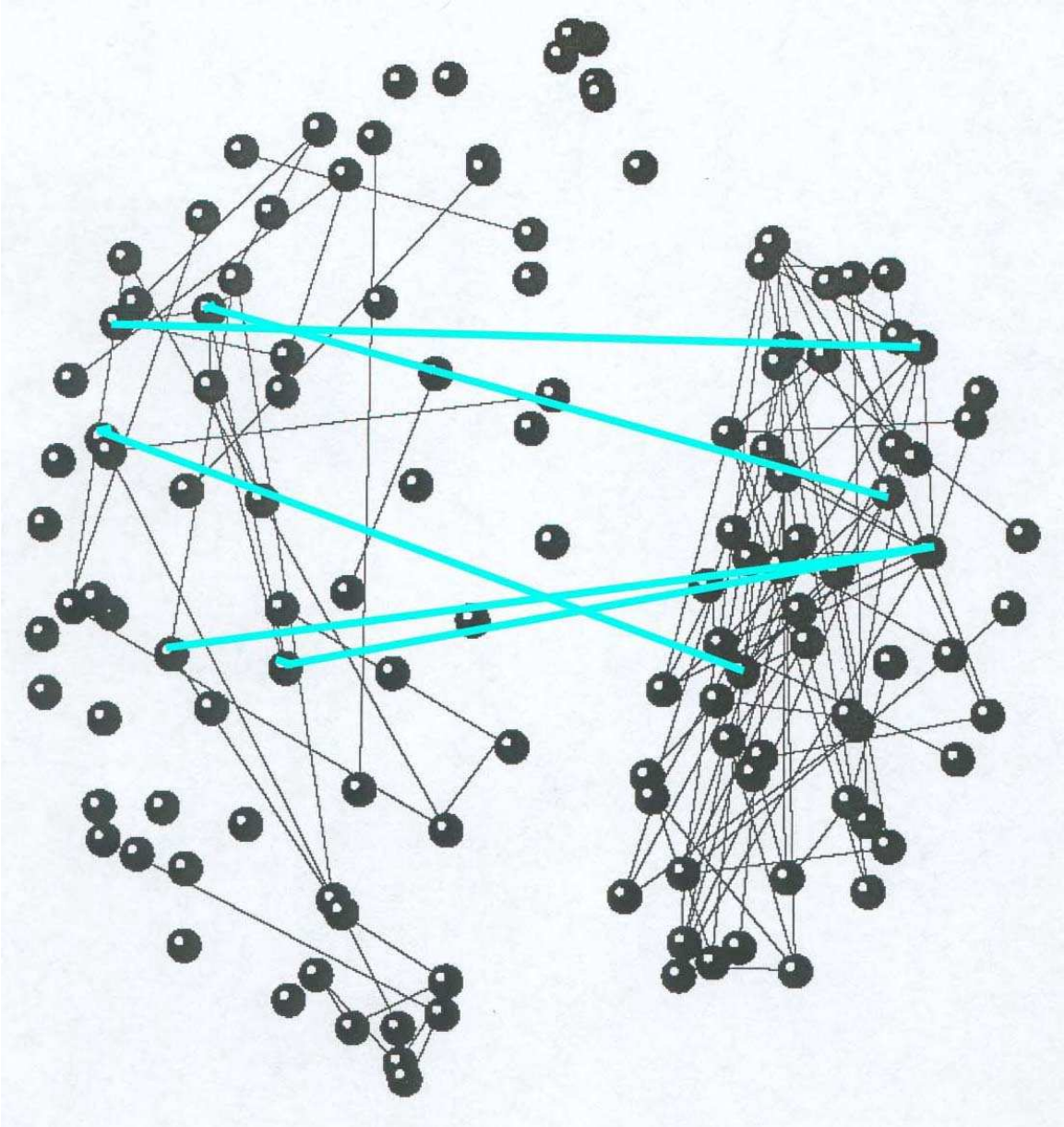

Figure 11: Nodes with the Smart Small World Policy Injection

A Third policy selection rule is that of Connectivity Fairness that extends the properties of the Smart Small World to reduce the path length for those nodes most disconnected to some distant colleague. The motivation for this rule is that success of a policy is dependant on keeping as many actors as possible in the system after each policy cycle, so connections are chosen between individuals so as to help keep

"disenfranchised" individuals in the system. The rule finds those nodes that are accessible to the most other nodes in the fewest number of jumps, and then connects them to each other on both sides of the graph. This rule is premised on fairness and does not approximate the lowest average path length recommended by Watts (1999) and Watts and Strogatz (1999). However, it does configure the network with the five selected connections so that the five longest paths in the network are as small as possible. The method for generating the candidates for connection using the fairness rule produces many ties and many choices that graphically do not seem like reasonable candidates for connectivity. Thus, the implementation of the Connectivity Fairness rule relies on the assumption that connections will generally not be broken and that information travels well between nodes in a graph. This can be OK in many instances, but as shown below Fairness explicitly targeted by this rule itself can be approximated or exceeded by other rules. Both Smart Small world and Smarter Small Worlds below, as we shall see, embed generous fairness properties under the condition that alienated connections have the potential to become productive connections if they can acquire more resources.

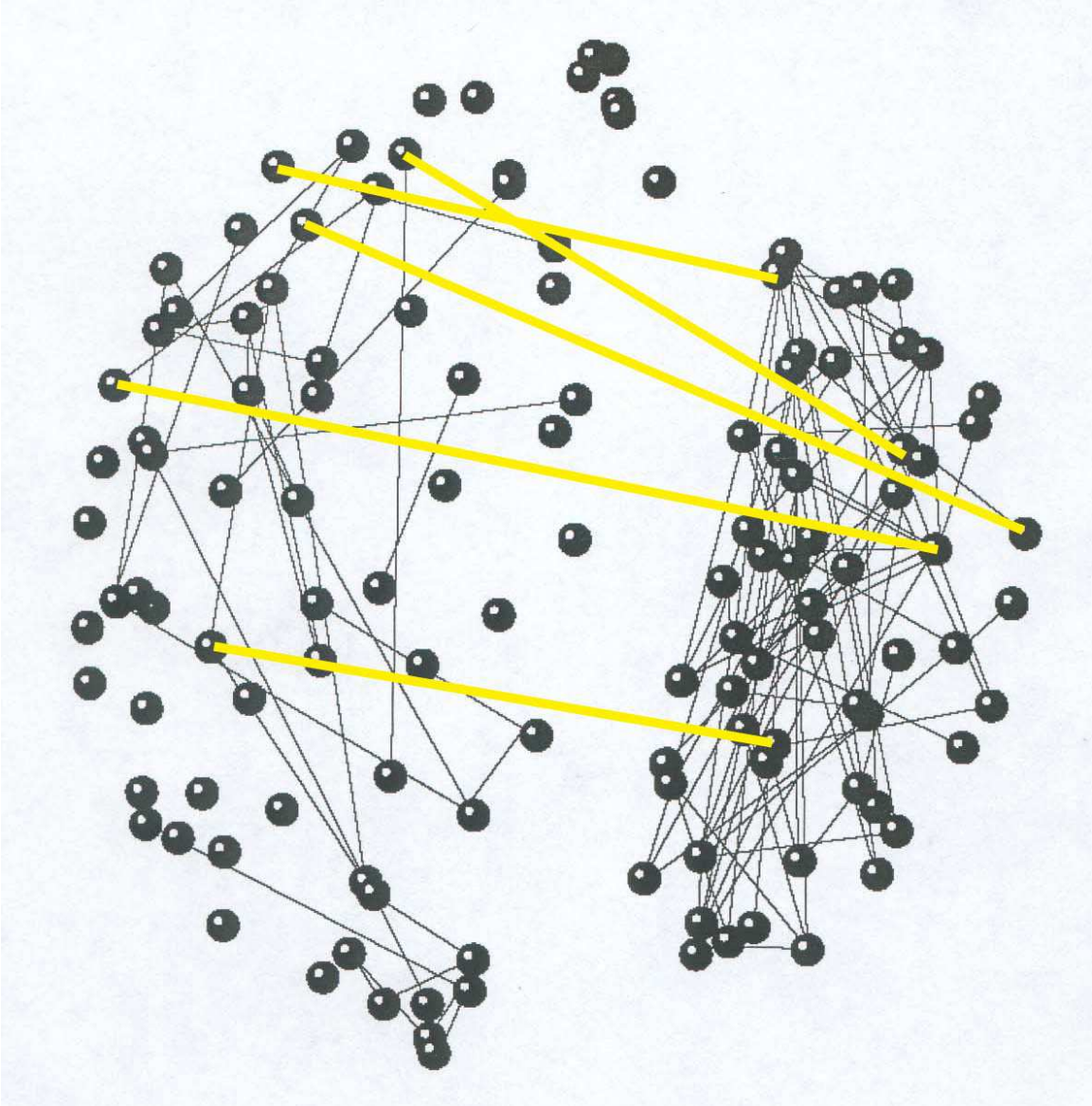

Figure 12: Connectivity Fairness Policy for the Castillo Dataset

A new *Smarter* Small World policy selection is based on the idea that the probabilities of connection are generally known and are tied to the number of connections from one node to another. It is applicable in environments where more information about the cellular automata rules is known and there is a large domain of connections that display the property of increasing returns. In a situation of increasing returns based on connectivity, different regimes of probability are set up so that a person within a certain

range of connections from themselves will have a certain probability and another person within another certain range of connections will have another probability. In a dense graph with wide variation in connectivity between individuals, there will be individuals who are just outside of the top probability regime by one connection, and if these individuals are used as connectivity candidates, their probability of connection will increase due to them moving into the top probability regime. This policy is different from that of the direct optimization or connection optimization because making a connection between the two highly connected nodes will not change their probability of connection in a situation of increasing returns, and it differs from the Smart Small World and Connectivity Fairness policies because this policy does not rely on structural characteristics beyond a single node in determining connection candidates.

It is noted that in multiple policy injection situations the use of the Smarter Small World policy has its limitations in terms of connectivity choice since the probability of connection is only determined before the first policy cycle. In the case of the Smarter Small World policy choice, the candidates for connection in the second and third policy cycles were determined through means of topography and not on potential output maximization. Thus, the candidates for the second and third policy cycles were chosen to be those nodes that were connected to the most other nodes in the system, similar (if not identical) to those nodes in the Connection Maximization Policy.

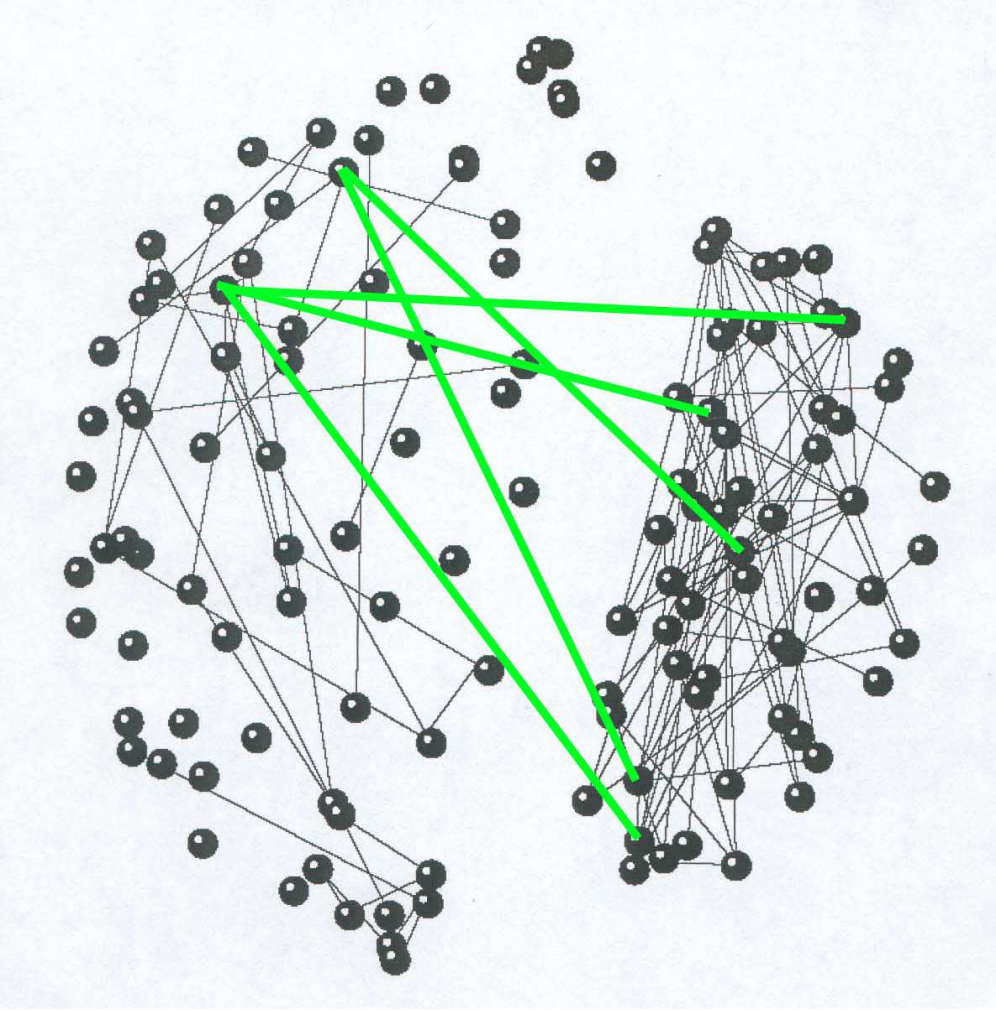

Figure 13: Smarter Small World Policy Connection for the Castillo Dataset

Each of the policy choices presented reflects a different fundamental attitude concerning policy theory, and each policy also can reflect a different amount of information concerning the graph of nodes. Some of the policies act on the system in a manner so as to minimize the distance between all individuals (a viewpoint similar to equity), while other policies desire to create connections that may be best for output increases but not for keeping many of the fringe nodes in the system (similar to the ideals of efficiency or rule utilitarianism)

Table 1: Policy choices based on the amount of knowledge concerning the system and also concerning attitudes concerning fairness in the system.

	Knowledge of the system of connections and nodes is *low*	**Knowledge of the system of connections and nodes is *high***
Fairness to all nodes in the system is *least important*	Direct Optimization, Smart Small World	Smarter Small World
Fairness to all nodes in the system is *important*	Connectivity Fairness, L/C minimizing Policy	Hybrid policy between the theoretical L/C minimizing policy and the Smarter Small World

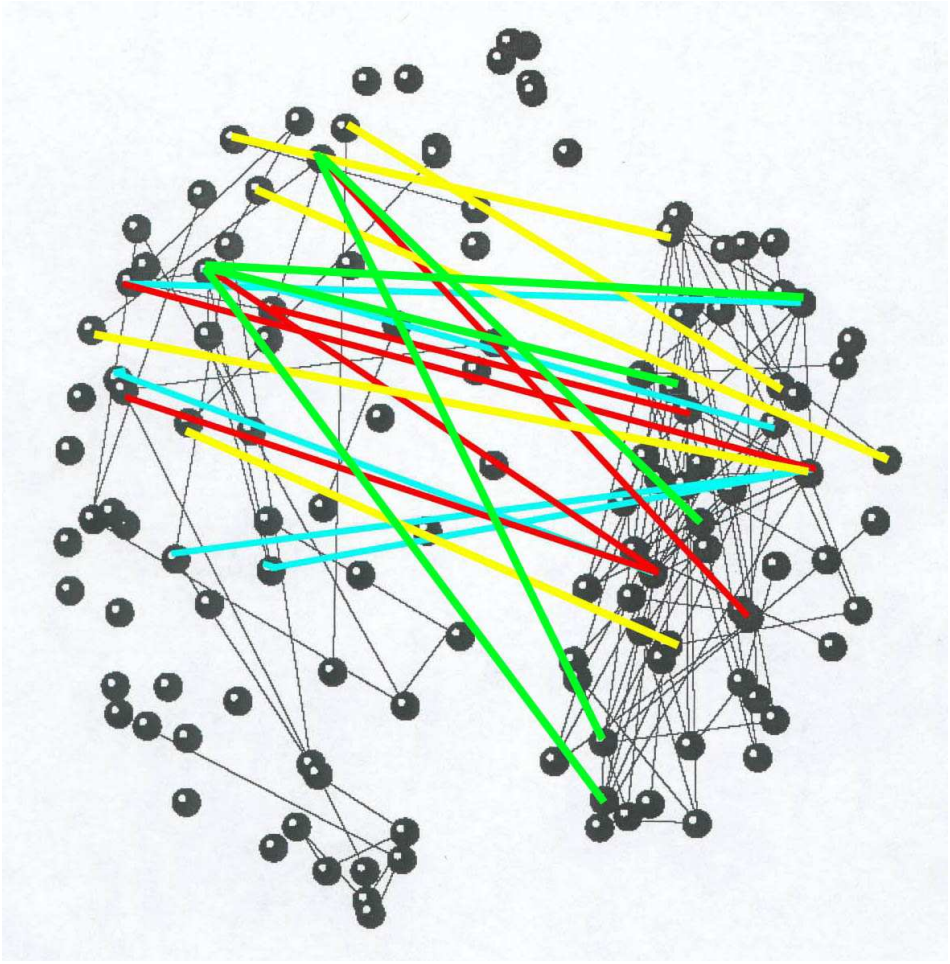

Figure 14: Graph of all the policy choices on the Castillo dataset

The Model

There are two main approaches toward the examination and the addition of dynamic rules to graphs of nodes in a network. One approach is to view the network as a single monolithic system whose dynamic is characterized by a highly nonlinear dynamic function (Dodds and Watts, 2004). This system-wide function is often quite complex and, as found in bottom up studies, seldom replicable from network to network; yet if associations are of rather uniform quality, such as infectious disease spread, single system dynamics can be quite predictive.

Another approach considers each node to be connected to other nodes where the properties of change for any given link obey rather simple rules. Network complexity will emerge from these very simple rules and a set of individual initial states. This is similar to the basic concept of complex dynamics in that initial conditions are important for finding the dynamics of a system. Connections may break in response to a network stimulus as other nodes may gain connections, but individual connections experience different results based on the initial position or state in which an agent exists. As connectivity ripples through the network and as each change produces its own ripple effect, these simple processes turn out highly diverse, complex outcomes for a system. This sort of dynamic is considered realistic for social network systems (Smith and Steven, 1999). In this type of model, to deduce universal properties from a given dynamic trajectory misses the point. It is the initial conditions (in the form of the network map with the initial connections injected into it) coupled with simple rules of evolution that determine the trajectory of the dynamics of the system. White (2003) charts network shapes by linking key diagnostics directly to different social concepts and

principles. With enough randomness, or probabilistic distinctions centered on how each individual association evolves or acts to affect its particular tasks, dynamic analyses not based on connection by connection characteristics can be a highly suspect endeavor that can mask core conceptual distinctions (White et al, 2003). This leads researchers to conclude that analyses that evolve as the product of reactions of each node in a positional context for completing a task.

Researchers have coined the phrase "cellular automata" (Wolfram, 1983) to underscore the independent choices of individuals or of particular connections as cells in a system that operate with relative autonomy but are connected by the inter-lacing of their autonomous actions in a network. So dynamics reduce to locating rather simple *dynamic* properties or 'cellular automata' rules from first principles that flow from the theory behind the network organization mapped, connection by connection. It turns out this is not difficult to do. Quite the contrary, rather than attempting singular system-wide operations, estimation of direct expectation into how individual associations might change under a probability distribution of choices or outcomes in a given period can be simulated numerous times to generate a picture of the various ways the network might evolve. This keeps the *dynamic* estimation structurally close to the theory that the network analyst employs to describe the network. A program to model the distribution of plausible dynamic outcomes easily can introduce a policy that alters the initial network and then track a distribution of outcomes that flow from that intervention where the dynamics are premised directly on the data of association characteristics that the analyst used to map and describe the network.

Cellular Automata Rules

The network of collaborations in joint ventures is used to simulate system effects from some simple properties of association (Wolfram, 1983).

First the output of a connection is weighted. Not all connections are equal because of nodal characteristics, connectivity characteristics, or both. In this case each connection is assigned a probability of success. Simulations here allow connections to succeed 10% of the time, 30 % of the time or 60% of the time. Second, the chance of output success is not static. From an initial assignment, success in one period improves the chance of success in the next: moving from 10% to 30% and 30% to 60% (where it peaks); or back down along the connection probabilities. Third, the number of raw number nodes and connections is also not static. Following the policy literature, a policy cycle should be more than one period and is chosen to be three periods in this model. Any connection that fails three times in sequence is broken; any node that fails to succeed in any collaboration over three periods exits the network. Any connection that succeeds three periods in row, adds a node, connected to both collaborators with a probability of success 0f 0.3 for each connection. Added sensitivity analysis was conducted over different probabilities and by cutting connections in half.

Finally, a comparison is done between increasing returns and constant returns success structures. In increasing returns, for those central nodes with a gross high number of collaborations, the probability of success on each connection is higher than nodes with few connections: or nodes with seven or more collaborators has a 60% chance of succeeding in any period with each connection, rising or falling from there as periods pass. Those nodes with between three and six connections had a 30% chance of success

in the first period, and those nodes with two or fewer connections had a 10% chance of success on each connection in the first period. (The code for this routine is found in the code section of the appendix)

One of the main motivating factors behind the use of the increasing returns model came from the work by Paul Krugman (Krugman, 1991), where he determined that economies migrated into a system of industrialized "cores" and agricultural "peripheries", which are not dissimilar to the hub and spoke model in social networks. Yet the theory is more flexible, or case specific. The industrialized cores tended to move into areas with higher demand, which is analogous to there being a higher probability of success. Likewise, in venture capital studies, there is something to be said of a firm that acts as a hub of cooperation to multiple other firms in an up and coming industry. If a company is receiving funding from the same venture capital firm as multiple other companies, then there is faith from multiple parties that their business strategy will be a success.

Here the critical and case specific structure of complementary economic activities can be activated. Unfortunately, in standard organizational theory analyses or stand alone regression model approximation approaches, there is too much averaging over connections despite the theory that critical positioning of a specific node drives success. In this sense the concerns of Krugman and Smith (1994) that strategic policy manipulation is unlikely as the key pivotal points (or hubs in the full context of the surrounding contingencies) are unobservable are eased considerably by the focus on how well a given connection functions on its task and how the given network operates as a network of independent actors with differentiated talents and prospects.

This is compared to a probability structure where the chance of success for any given collaboration was randomly distributed over the map, and there were no positive returns to scale in this model. (The code for this routine is found in the code section of the appendix) Simulations run 100 times to create a histogram of outcome distributions and to help determine the statistics for the simulation.

Table 2: Connection probability determinations as a function of each probability regime.

Probability Regime	Values
Increasing Returns to Scale	If a node has two or fewer connections to other nodes, it gets a probability of 0.1. If a node has more than two connections to other nodes and less than six connections to other nodes, its connections get a probability of 0.3. If a node has more than six connections to other nodes, its connections get a probability of 0.6.
Constant Returns to Scale	Connection values are assigned the numbers 0.1, 0.3, and 0.6 at random

Joint ventures are funded (through means of adding connections) in the initial cycle, but the policy injection does not stop before the first cycle. At the end of a policy cycle, three periods in these simulations, a new network map is constructed. Policy-makers re-map the network at that time and the policy rule is redeployed at the start of period four to sponsor another group of joint ventures (through means of adding additional connections into the system). The policy analysis, implementation and evaluation process repeats again at the end of period six and a final set of collaborations is funded. Total output cumulative up to the end of period nine is compared as well as the overall shape and character of the final emergent network.

Policy Simulation Results

Table 3 shows summary information on the immediate period 1 output for constant returns, while Table 4 gives the summary information for the immediate period 1 output for increasing returns. Note that since there is a bifurcation in the nodes in the baseline, the L/C ratio is not reported. For constant returns, there was little variability in the Immediate period one output since the connection probability was normalized. The L/C ratio of the four policies was interesting because the policy that was intended in reducing the L/C ratio actually had the highest ratio, while the Direct Optimization Policy, with its Hub and Spoke model of connecting nodes, had the lowest L/C ratio.

The final output of the system was to within one standard deviation of each other for all policies except for that of the Smarter Small World, which greatly overpowered the other policies. Likewise, in the calculation of the final nodes in the graph after the addition of nodes through successful connections, the Smarter Small World policy greatly outperformed all of the other policies.

Table 3: Constant Returns Period 1 Summary Output

	Immediate Period One Output	L/C Ratio	Total Output (mean)	Final Number of Nodes (mean)
Baseline	68.0	Un-calculable	59.0	175
Max Connection	68.6	4.3599	60.4	179
Smart Small World	68.5	4.5068	69.0	189
Connectivity Fairness	68.4	5.8359	62.3	179
Smarter Small World	65.4	4.7162	77.5	206

For the increasing returns, the performance of the Smarter Small World is accentuated since the connections were chosen as a function of how they would perform with the addition of additional nodes into the system. As Table 4 shows, from the immediate period one output forward in the policy cycle, the Smarter Small World policy outperformed the other policies.

Table 4: Increasing Returns Period 1 Summary Output

	Immediate Period One Output	L/C Ratio	Total Output (mean)	Final Number of Nodes (mean)
Baseline	71.5	Un-calculable	63.9	194
Max Connection	74.5	4.3599	69.4	202
Smart Small World	75.9	4.5068	76.7	210
Connectivity Fairness	73.9	5.8359	73.5	211
Smarter Small World	91.41	4.7162	86.74	228

For the output of the system after three policy cycles (or after 9 periods and three policy injections) in the constant returns regime, there are many topographic items of interest. As is seen in Table 5, even with no bridge connections the baseline had the most number of reachable nodes. This is an anomaly of the simulation which just came out more strongly over the 100 test simulations than did the other data since when the data was re-run the baseline came out much worse in the average number of reachable nodes. What is interesting is that the Max Connection, with its use of the hub and spoke model for policy choice, was significantly more useful in moving to accessible nodes than the other policy choices behaved.

Table 5: Constant Returns Period 9 Summary Table

Policy	Bridge Connections Made over Nine Policy Cycles	Average Total Jumps to Available Nodes	Average Number of Left Out Nodes	Average Number of Reachable Nodes
Baseline	0	317.579	46.63	167.364
Max Connection	15	213.79	45.211	153.79
Max Fairness	15	310.05	45.12	145.32
Smart Small World	15	332.26	64.09	131.91
Smarter Smart Small World	15	490.00	44.90	161.00

In the case of increasing returns, as is seen in Table 6, it is noted that the smarter small world policy behaved quite well in generating the largest number of reachable nodes, as is expected since the function of the Smarter Small World policy is to maximize the probability for connection in the system. Something that is not so promising about the smarter small world policy is that the system's connections are not as central or as useful as those in the Max Fairness simulation because the choices for connection were not chosen due to geographic considerations but were instead chosen based on just their raw number of connections to other nodes. The Smart Small World, while generating a large amount of output, did not have a large number of reachable nodes in either the constant or increasing returns situations. The main reason for this is that because the policy connects those nodes that are not the hub of a hub and spoke but are the most important spoke, if the connection to the hub stays then the output will be enormous and the accessibility will be basically the same as the max connection policy choice, but if the connection to the hub breaks the output of the system will be mildly reduced (since the output is a measure of all connections and has nothing to do with the topography of the

nodes) while the reachability of the system will be greatly reduced since the connection to the hub is of extreme importance to the systems reachability.

Table 6: Increasing Returns Period 9 Summary Table

Policy	Bridge Connections Made over Nine Policy Cycles	Average Total Jumps to Available Nodes	Average Number of Left Out Nodes	Average Number of Reachable Nodes
Baseline	0	269.103	51.1752	142.82
Max Connection	15	514.08	97.504	124.09
Max Fairness	15	220.56	54.44	136.55
Smart Small World	15	336.12	79.18	113.13
Smarter Smart Small World	15	490.03	52.26	175.00

Data Conclusions

There is a risk associated with each of the policy choices in working with the Castillo dataset because there is a strong dependency on what knowledge one has about a system when making connections. While policies that focus on topography are good because they can minimize the distance to nodes, they have their downfall in situations when a policy choice leads to an unstable nodal connection. The upside to the policy of the Smarter Small World is that it has a direct correlation between the use of the policy and the raising of output in the system over all nodes. The downside to this policy choice is that it requires a high level of knowledge throughout the system, and it is not so plausible in the real world to find those nodes whose output potential is just on the brink of success. The amount of back end research that would go into finding those nodes "on the brink" of success may be more trouble than it is worth.

A large caveat to the use of the Smarter Small World policy choice is that the policy basically completely forgoes the use of the static network graph and instead relies on the use of future potential in determining network candidates from a well defined initial condition. While it is clear from the results that there is power in the knowledge of which nodes are at the greatest potential for growth of potential output, this knowledge is probably impossible to ascertain without much fieldwork and modeling of many variables concerning individual nodes and their connections.

Prelude to Chapters 3 and 4

The data that was examined in the previous chapter pertained to a system where the goal of policy was to stimulate innovation and accelerate output among the sum of all the companies. The policy in the remaining two chapters is similar in that it also strives for stimulation, but the mechanism for growth will now be placed in an entirely different framework. In the final chapters the goal of policy is to help motivate a research environment.

There is much interest in stimulating research environments because is widely believed that an academic research environment is a strong facet of the national innovation system of a country (Crow and Bozeman, 2002). Also, as competition for finding the "best" researchers in a field is intensifying between universities and between states, there is a new interest in what it means for a researcher to be the *best* in their respected field. Some would argue that the best researcher in a field is someone who publishes the most number of articles or has their articles cited more times than other researchers, while others would argue that a best researcher in a field would be a person who brought in the most funding for a lab (Tijssen, 2002). These are good measures of strength of a researcher, but the paradigm of research output potential could be shifted from these monolithic measures to an automaton-like system where a researcher's peers and collaborators are mapped together so that relevance is not only determined by output but by their position among their peers.

In a network of researchers who have co-authored journal articles, the scope of the problem for advancing research through funding collaborative studies goes back to

the two ideas from the Castillo dataset of minimizing the distance that information has to travel and maximizing the benefit (in this case, the research output) for the cost of funding a joint research venture. What is nice about social networks in this framework is that there is an already existing and expanding set of tools that are usable in figuring out an individual's research capacity, so more accurate weights for individuals and connections are possible.

Current research on forecasting future output by academic researchers mainly focuses on either bibliometric studies (Murray (2003); Hicks et al (2000)) or career path studies (Dietz, 2004). This is a useful piece of information, but its utility could be further maximized if they were also placed in the context of a research network environment. For instance, a researcher might not have a high score on some traditional bibliometric measures because he or she has not published that many articles or maybe has not had their articles cited or are not at a good point in their career path. A network analysis might show that their position is integral to the system because they bridge two important disciplines or create a communication link between two researchers that are productive by traditional researcher studies. In a study of who important researchers are in a field, it might be just as important to cite central nodes of a graph of researchers or characters who bridge as much as it is to find those researchers who are the most productive by existing standards.

The first of the two forthcoming chapters deals with how the social network analysis of the second chapter could be applied to a set of academic researchers in an interdisciplinary field. Policies can be applied to mappings of co-authorship alliances in a manner similar to the venture capital map from a similar set of assumptions. The final

chapter of this thesis explains a procedure to determine analytically and computationally what the important groups are in a research environment through means of a principal component analysis and also who the important researchers are in that environment based on their output in those newly found important groups.

Chapter 3: Data Set 2 and Second Analysis Using Social Networks

Background

In science and technology, the jump from basic research to applied technology and end user products can be a difficult task. This is partly because connections are needed between groups of individual researchers that normally do not work together. It is not necessarily that these different research groups are incapable of working together but that their research interests are just in different areas and they are not exposed to each other's research interests. If interdisciplinary joint research ventures are funded there could be quicker research milestones that could lead to a quicker applied technology. This is supported by the research of social scientists working in the fields of innovation policy. (Salter and Ben Martin, 2001) Also, there is research on how innovation can move more quickly through collaborative strategies. (Aghion and Howitt, 1992)

Science and Technology policy is not just about finding means to increase productivity between researchers, but it can also be used as a tool of economic development. For example, if some communities have their growth pole as a business that relies on technology for sustainability, it is important for the technology to proceed as quickly as possible. Nowhere is this scenario more realistic than in the paper mill industry in the United States.

For paper mills in America, there is a huge administrative concern because many mills are operating in noncompliance with future and even present environmental standards. Also, there is much waste that comes from the process of turning a tree to

paper, as only certain parts of a tree can be used for processing into paper. The rest of the biomass (usually in the form of black liquor) is not usable for other standard wood products, but research in a new process called black liquor gasification (henceforth referred to as BLG) has shown promise for turning this biomass into electricity in a gas turbine combined cycle. Upon further processing, the BLG process can convert black liquor into pure transportation "syngases" such as Fischer-Tropsch liquids or hydrogen. So there is both an economic and environmental push for processing biomass fuels from paper mills.

There are several reasons why this combined research is vital to the American paper industry. Certain facilities in Europe that have developed early BLG processing modules have found that there are large up front costs associated with developing the technology, so it is paramount that if the United States is serious about this technology they should work on keeping sunk costs as low as possible. Also, from an economic development perspective, this technology may save several paper mill communities because the local mill may be the largest employer and is the growth pole around which other agglomeration economies gather in certain local communities. If these mills were shut down through environmental noncompliance or through economic competition with global providers it could be catastrophic to these regions.

Data

From bibiometric analyses, co-authorship alliances can be mapped through use of such programs as VantagePoint. Within each of these alliances, certain groups of keywords in publications can be isolated and grouped together into larger topics by experts in the field. Hence, a simplified graph of each of the main groupings (or super-topics) of a research group can be found and mapped. Three of the main areas of research in BLG processing are that of biomass processing (focused in the areas of gas turbines, pyrolysis, and waste liquor utilization), Reaction Kinetics (focused on mass transfer, high pressure effects, and CO), and Spent Liquors (focused on the modeling of the process and the characterization of them as they are being processed). As a note, each of these three groups of research alliances has some overlap into the other two research groups, but it is not significant as their connections in their own group due to the keywords used in their research or the amount of times that a researcher published with a researcher in one group versus another group. A graphical representation of the groups of researchers is found in Figure 15.

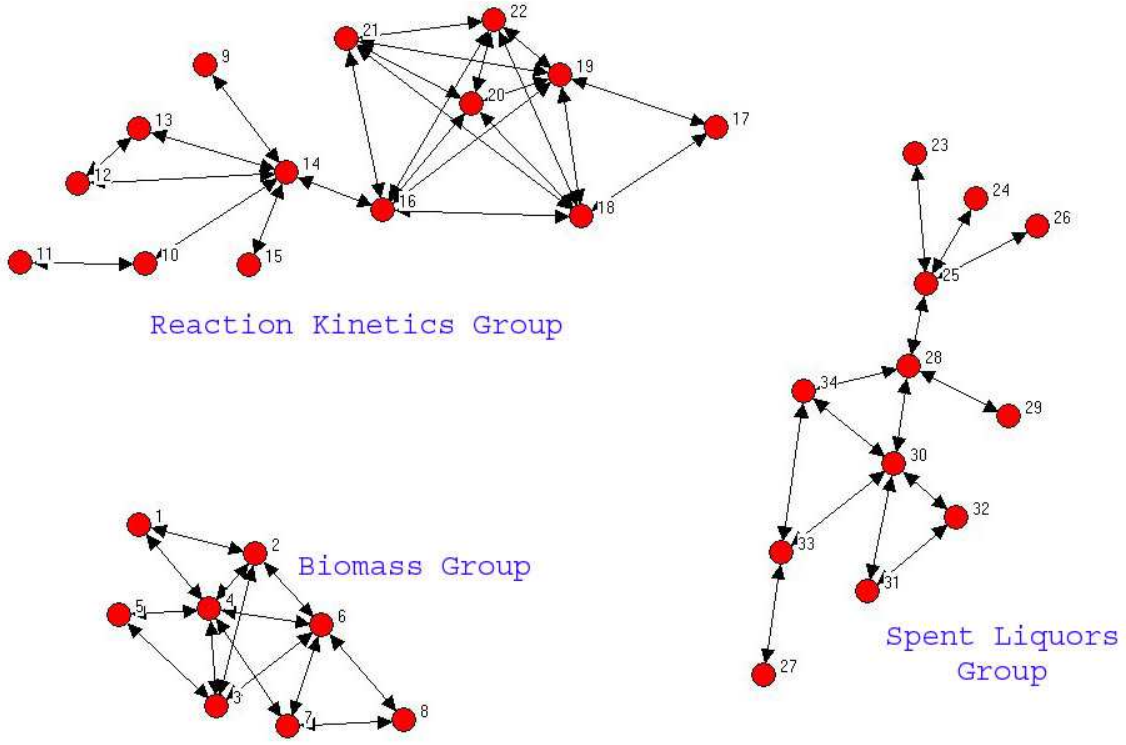

Figure 15: Clusters of Researchers in Biomass, Reaction Kinetics, and Spent Liquors. #Figure Generated with UCINET, NetDraw, and Adobe Photoshop

Each of the numbered nodes in figure 15 represents an individual researcher, while a connection between two nodes represents a co-authorship. Some researchers, due to their general productivity or due to other factors, have more co-authorships connections than do some other researchers. This is due to a myriad of factors, including but not limited to author productivity, author prestige, and author openness to work with other researchers.

Caveat in Following Analysis

There is a forewarning to the analysis being used in this paper because graphs of these sizes are usually dealt with "by hand" because, in lieu of computational techniques, certain experts on each of these research areas might have a better idea as to what connections should be generated between researchers in different groups. Usually, highly analytic or computational techniques have their most power in systems of graphs with many nodes (at least on the scale of at least a hundred nodes as in the Castillo dataset) and are hard to implement on smaller datasets with fewer nodes and fewer policy choices.

Graphical Data

As was found in the Castillo dataset, there is 100% connectivity within each of the three research groups, meaning that one researcher can traverse to all other researchers within their topical group. Each of the subgroups is represented by a dense connectivity graph as is explained by the number of "jumps" needed to traverse a sub-topic. Some topics such as BioMass can be traversed in as few as three jumps, while it only takes as many as five jumps to traverse the Reaction Kinetics and Spent Liquors groups.

The number of connections between nodes as a function of all potential connections determines total connectivity. For example, if every node were connected to every other node in the system, there would be a total connectivity of 1.0 in the system. Also, the total connectivity is also equal to the average connectivity for all nodes in the system. It is noted that nodes in these graphs are connected to themselves, and self-connections are counted in the average number of connections, most number of connections, and least number of connections.

The total graph of all the connections has a trivial density of 9.3% and a non-trivial density of 9.4%. Since the graphs are not that large, it was determined that three funded connections between researchers within each of the three groups would be the ideal policy. This would raise both the trivial and non-trivial density by 0.3%.

Table 7: Summary Data on the three research groups.

	BioMass Group	**Reaction Kinetics Group**	**Spent Liquors Group**
Total Connectivity	0.5625	0.3265	0.3265
Fewest Jumps needed to Traverse System/Graph	3	3	3
Most Jumps Needed to Traverse System/Graph	4	5	5
Average number of Connections	4.5	4.57	3.33
Most Number of Connections	7	7	6
Least Number of Connections	3	2	2

#Values generated using Matlab

Policy and Numerical Analysis

As was said before in the Castillo data set, the main goal of policy in this simulation is to maximize the benefit for the cost. In this model, the number of successful connections will correlate directly with the amount of successful research to come from the system. As opposed to a correlation of a successful connection with a successful venture, as how the model works in the Castillo dataset, here a successful venture will equate to a successful research collaboration that results in a publication. It is trivial but noted that publications help to bring the research from the basic stages to the advanced development and final product stages.

Like the Castillo model, the model for the BLG system will be based upon the principle of the policy cycle, success breeding additional nodes in the system and failure leading to broken connections, and probability of connection between two researchers is not always easily determined.

In this analysis, there will be three probability regimes based upon three different philosophies of research output. While the Castillo dataset had two probability structures based on constant and increasing returns to scale (which are denoted here as Prob1 and Prob3), in this more academic model it is important for there to be a probability model with decreasing returns to scale. The first probability regime (denoted Prob1 or called increasing returns to scale) will be based on the premise that larger number of connections between a node and other nodes will correlate with a higher probability of connection between those nodes. This is the “success breeds success” principle. The second probability regime (denoted Prob2 or called decreasing returns to scale) is based on the principle of divided time, meaning that if a researcher is working on a project with

only one other coauthor that there is a greater chance of those researchers working together again as opposed to a researcher working with multiple people and not being able to spend as much time with them and lowering the probability of a co-authorship. In other words, a researcher who is only working with one or a few other collaborators will have much more time for correspondence with them and will have a higher probability of creating research outputs with them. The final probability regime (denoted Prob3 or called constant returns to scale) is that all connections between researchers are stochastically determined and that there is no basis for determining probabilities of connections between researchers.

There are three different policies that will be applied to this system. The first policy (called P1 in this section) is that of a greatest output of connections made. This is identical to the Max Connection policy in the examination of the Castillo dataset and is based on the idea that researchers that are more highly connected will have a greater probability of future output than those researchers that work with one or a few other researchers. This may end up alienating certain fringe researchers after one policy cycle, but it will create even stronger ties for those nodes that are the most stable and those that are most likely to produce output in the system.

The second policy (called P2 in this section) to be implemented will be that of a fairness rule where the average distance between all nodes will be minimized (this policy is identical to the Fairness Rule of the Castillo section). This policy can sometimes lead to fringe/non-central nodes being connected or nodes that analytically are chosen to be connected while graphically may seem to have no reason whatsoever for being a candidate for connection.

A third policy alternative (called P3 in this section) would be to institute connections based on nodes that have the highest number of connections two steps away from themselves. (This policy is identical to the Smart Small World Policy from the Castillo section) This rule is based on the idea that if connections are going to be broken but not at a high rate it might be advantageous to look for nodes that are central to the system but are not the most central. Policy P3 could be considered a compromise between policies P1 and P2, and it can have results that are a large departure from either of them. Policy P3 is based on the idea that researchers, their collaborators, and their collaborator's collaborators would be as far reaching as possible for information transmission. An example of this would be that a researcher could ask his or her collaborators about a certain topic. The researcher's collaborators may know of another cohort that is working on a project, but that is probably the extent to which the information can travel from one researcher to another. The output maximization policy in terms of the Smarter Small World was not implemented on this dataset because there was no uniqueness to the policy candidates versus the candidates of the other policies.

Connection Choices based on Policy

For policy P1 (greatest output of connections made), the recommended connections are between the nodes 4 and 16, 16 and 30, and 4 and 30. The connections look as the following:

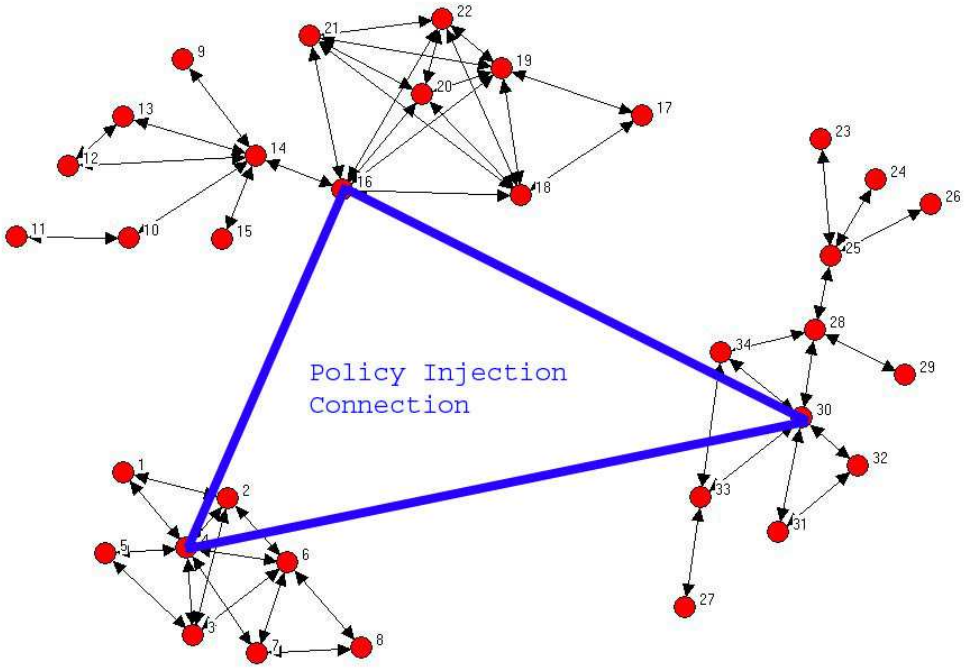

Figure 16: Policy P1 connections
Figure Generated with Matlab, UCINET, NetDraw, and Adobe Photoshop

For policy P2 (minimize average distance between all nodes and maximize fairness for all nodes in the system), the recommended connections are between the nodes 6 and 16, 6 and 28, and 16 and 28. The connections look as the following:

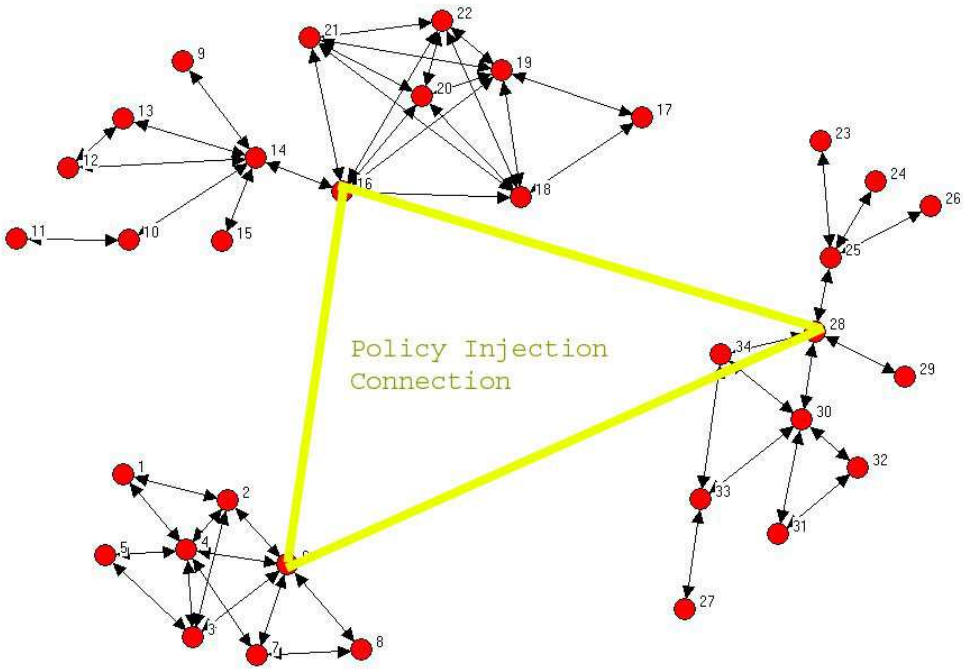

Figure 17: Policy P2 connections
Figure Generated with Matlab, UCINET, NetDraw, and Adobe Photoshop

For policy P3 (highest number of connections two steps away from themselves), the recommended connections are between the nodes 3 and 14, 14 and 28, and 3 and 28. The connections look as the following:

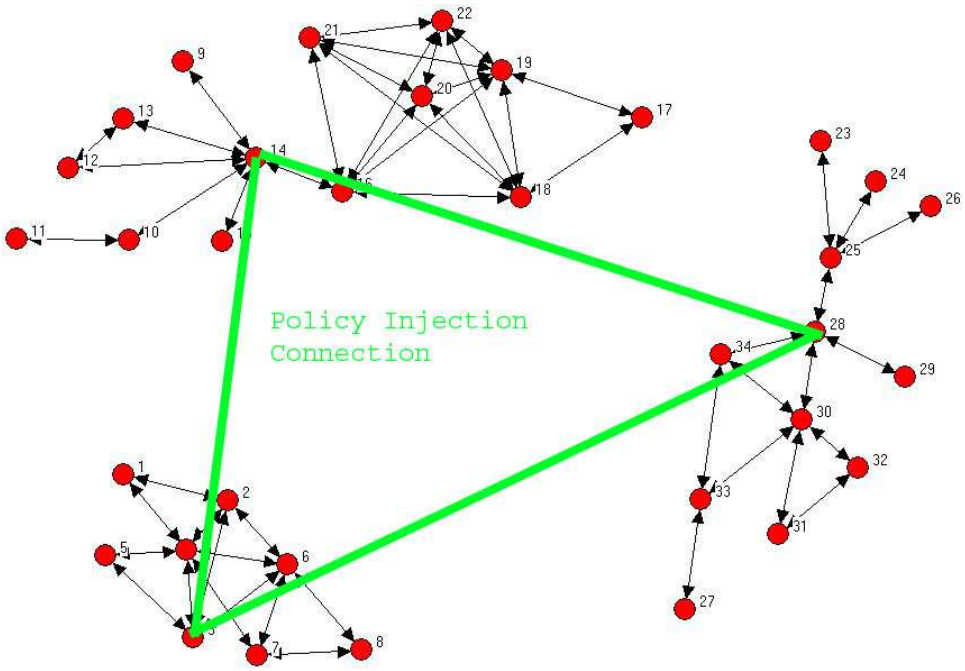

Figure 18: Policy P3 connections
Figure Generated with Matlab, UCINET, NetDraw, and Adobe Photoshop

For comparison, all policies are placed on the same image so that the discrepancy between policies is evident.

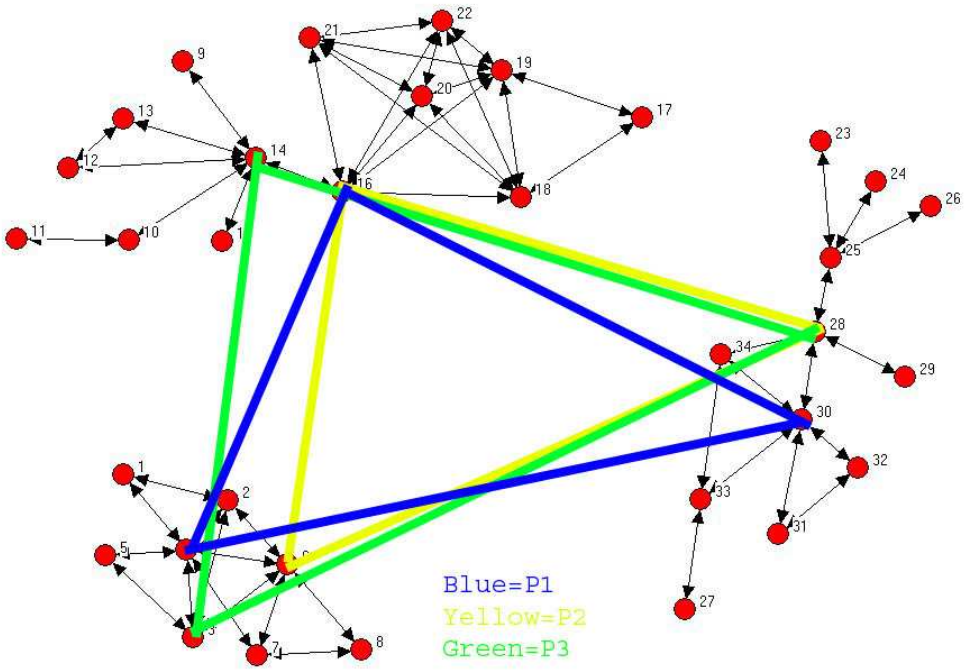

Figure 19: Policy P1, P2, and P3 connections
Figure Generated with Matlab, UCINET, NetDraw, and Adobe Photoshop

Model

Initially, all connections in the system are assigned a probability based on one of the three probability regimes mentioned in the Numerical Analysis of Policy section. The three probability regime values are explained in Table 8. The connection probability values are based on three probability levels, each representing a likelihood of connection. The 0.1 probability is designed so that most of the time it will not lead to a connection between nodes. The 0.3 probability value is based on the idea that a connection is not likely to connect but is not as unlikely to connect as the 0.1 connection probability. The 0.6 probability is the most likely to produce a connection but is still not definite (1.0 is the only probability where an output can be guaranteed), as in many research environments there are many reasons why researchers are likely to produce an output yet still do not.

Table 8: Connection probability determinations as a function of each probability regime

Probability Regime	Values
Prob1 or Increasing Returns to Scale	If a node has two or fewer connections to other nodes, it gets a probability of 0.1. If a node has more than two connections to other nodes and less than five connections to other nodes, its connections get a probability of 0.3. If a node has more than five connections to other nodes, its connections get a probability of 0.6.
Prob2 or Decreasing Returns to Scale	If a node has more than five connections to other nodes, its connections get a probability of 0.1. If a node has more than two connections to other nodes and less than five connections to other nodes, its connections get a probability of 0.3. If a node has two or fewer connections to other nodes, it gets a probability of 0.6.
Prob3 or Constant Returns to Scale	Connection values are assigned the numbers 0.1, 0.3, and 0.6 at random

The model for this system's policy cycle is identical to that of the Castillo system, except that this system is only taken to one policy cycle with three periods. The main reason why the system was not taken to two more policy cycles as in the Castillo system is that there are few choices of connection in the system, and the candidates for connection for subsequent policy cycles would end up being the same as the initial choices in the other two policy candidates, so all three policy choices over nine periods and three policy cycles would end up being almost identical to one another.

Results

After running the simulation on the graph network one hundred times for each probability regime and each policy connection choice (making for a total of twelve different simulation scenarios), the averages for total output and total coverage are listed in Table 3. The full results for every policy in every probability regime are listed in the appendix. For Probability regime 1, policy P3 has the largest output and is beyond the average of the other two policies plus one standard deviation. For probability regime 2, policy P3 again has the largest average but is not past the averages of the other two policies plus their standard deviation. Still, the upper tail of the P3 policy's output histogram extends beyond any of the other policies. For probability regime 3, there is little statistical significance to policy P2 being the greatest number because all policies are within one standard deviation of each other. It is noted that the output of policy P3 plus its standard deviation is the largest absolute output of the three policies.

In terms of coverage of the system, there is nothing that can be drawn from any of the results except that the policies are comparable in keeping the coverage of the system the same. This is likely because the graph is relatively dense in the trivial and non-trivial sense.

Table 9: Summary Results across all probability regimes and policies. Bold numbers represent the number that is the largest across a row.

	Baseline	P1	P2	P3
Prob1 Total Output	27.32262	31.93716	34.20074	**36.89587**
Prob2 Total Output	15.79622	20.12153	20.85871	**22.94813**
Prob3 Total Output	14.68157	15.23329	**15.33438**	15.01123
Prob1 Coverage	0.252244	0.724516	0.742951	**0.75837**
Prob2 Coverage	0.20109	0.464699	**0.473246**	0.443444
Prob3 Coverage	0.21843	0.534834	**0.564712**	0.524189

Numbers generated in Matlab

Discussion

There is a compromise between a policy of total fairness and a policy of maximizing connections. Especially in situations where there is a stochastic system of connection probability between researchers, it seems advisable to take this middle ground approach. A researcher's academic social network may not extend much further than two connections from themselves, so it is also advisable from a practical standpoint to find those researchers that have the highest number of research connections two steps from themselves.

This dataset has its advantages over the Castillo dataset of section 2 of this thesis in that there is more information concerning individual nodes. Also, since the nodes reflect actual researchers, a follow up study could be used to determine if individuals who are candidates for collaborative funding would be willing to partake in such a venture.

Chapter 4: Finding the appropriate way of finding the Maps for a Model

Introduction

For the final analysis, a set of components was chosen by means of a bibiometric analysis in the Black Liquor Gasification (BLG) dataset. This is different from the second data analysis set using BLG data because there were no assumptions as to which data grouping will be dominant in the end, and an unbiased factor analysis was used to determine which groups of variables are the ones that are left after variable reduction. In the second analysis, it was assumed that Kinetics, Biomass, and Spent Liquors were just the most important groups of research.

Data Reduction

Searches pertaining to Black Liquor Gasification were again done to populate a database of several hundred articles and almost a thousand key terms. Upon interviewing experts in the paper industry (Research was conducted by Dr. Michael Farmer with various researchers at the Institute for Paper Science and Technology on the campus of the Georgia Institute of Technology.), a set of thirty-four expert super-categories was produced. Each of these 34 categories of terms was then placed in a correlation matrix with each of the keywords from these articles. If a keyword matched with one of the expert terms, a value of 1 was placed in the correlation matrix for that element. If the keyword did not match with the expert term, then a value of zero was placed in the correlation matrix for that element.

Table 10: A list of the 34 Keywords that were determined by experts in the BLG industry

Gasification	Recovery	Thermodynamics
Combustion	Waste liquor utilization	Chars
Spent liquors	Evaporators	Power Generation
Pyrolysis	Kraft pulp	Environmental protection
Kraft process	Chemical Reactors	Industrial furnaces
Paper and pulp mills	Fluidized-Bed	Temperature
Kinetics	Gas Turbines	Bleaching
Biomass	Pressurization	Cost effectiveness
Thermal effects	Swelling	Carbon Monoxide
Carbon Dioxide	Paper and pulp industry	Sulfur dioxides
Nitrogen Oxides	Combined-Cycle-Power-Plants	Black liquors
Lignin		

Next, each of the columns in the correlation matrix was summed to determine a first approximation of the strength of some of the keywords. Many keywords only appeared a few times in the correlation matrix, so they were removed because they would not be statistically significant for any principal component analyses/factor analyses.

Also, some key terms that were generated from the expert list were combined due to their relative similarity to each other. For example, “Kraft Process” and “Kraft Pulp” were combined into a new variable called “Kraft Total”. Afterwards, the original set of 34 variables reduced down to 10 variables.

Table 11: Each of the Remaining Keywords with their number of occurrences (Note that the three bolded terms were combined from two keywords into one)

Component	Number of Times a Sub-Keyword Correlates with a key term
Gasification	116
Combustion	80
Spent liquors	53
Pyrolysis	48
Kinetics	37
Biomass	26
Thermal effects	25
Kraft Total	61
Paper and Pulp Total	56
Combined Cycle Power Plants with Gas Turbines	32

From this reduced set of variables, a principal components analysis was run on the data and the eigenvalues, eigenvectors, and scores for the variable analysis were reported. Finally, each of the observations (which in this case are the sub keywords) was grouped into each of the ten categories, and a histogram of the total number of matches with each of the sub-keywords is generated. The code used to generate the results is found in the appendix.

For this already reduced ten-variable system, the principal component analysis reduced the dataset to a seven variable system.

Table 12: Eigenvalue and difference statistics for the ten-variable system analysis

Component	Eigenvalue	Difference	Proportion	Cumulative
Gassification	1.7568	0.11897	0.1757	0.1757
Combustion	1.63783	0.41769	0.1638	0.3395
Spent Liquors	1.22014	0.18297	0.122	0.4615
Pyrolysis	1.03716	0.10277	0.1037	0.5652
Kinetics	0.9344	0.12139	0.0934	0.6586
Biomass	0.81301	0.02982	0.0813	0.7399
Thermal Affects	0.78318	0.07051	0.0783	0.8183
Kraft Total	0.71267	0.07912	0.0713	0.8895
Paper and Pulp Total	0.63355	0.16229	0.0634	0.9529
Combined Cycle Power Plants with Gas Turbines	0.47126	.	0.0471	1

Table 13: Eigenvectors for the ten-variable system.

Variable Eigenvector	Gasification	Combustion	Spent liquors	Pyrolysis	Kinetics	Biomass	Thermal effects	Kraft Total	Paper and Pulp Total	Combined Cycle Power Plants with Gas Turbines
Gassification	0.55137	0.12715	-0.04956	-0.11988	0.29802	-0.38445	0.13685	-0.2187	-0.01971	-0.59953
Combustion	0.10749	-0.13378	-0.04728	0.89243	0.02922	0.05044	0.28902	0.06035	-0.2767	-0.06883
Spent Liquors	0.38784	-0.29759	-0.45207	0.03064	0.29582	0.01317	0.08922	-0.07004	0.45923	0.49424
Pyrolysis	0.3747	-0.15716	0.41891	0.16717	-0.18659	-0.01124	-0.56675	-0.47653	-0.07946	0.20474
Kinetics	0.3	-0.35684	0.12711	-0.26022	0.23346	0.68247	0.0166	0.23462	-0.31676	-0.15122
Biomass	0.4227	0.22361	-0.10658	0.10719	-0.47149	0.04465	-0.26811	0.60528	0.26269	-0.1301
Thermal Affects	0.1795	-0.08834	0.68784	-0.09519	-0.01284	-0.26503	0.48674	0.2714	0.13626	0.27972
Kraft Total	-0.11246	0.34285	0.24327	0.21369	0.70993	0.01633	-0.39751	0.279	0.14898	0.05146
Paper and Pulp Total	0.07609	0.51012	0.14927	0.08812	-0.08157	0.55632	0.30589	-0.37855	0.38475	-0.05384
Combined Cycle Power Plants with Gas Turbines	0.27255	0.5383	-0.18264	-0.13641	0.01958	-0.03989	0.08894	0.01848	-0.58966	0.47545

Table 14: Scoring for the ten-variable system

Scoring of Variables	Gasification	Combustion	Spent liquors	Pyrolysis	Kinetics	Biomass	Thermal effects	Kraft Total	Paper and Pulp Total	Combined Cycle Power Plants with Gas Turbines
Gassification	0.55137	0.12715	-0.04956	-0.11988	0.29802	-0.38445	0.13685	-0.2187	-0.01971	-0.59953
Combustion	0.10749	-0.13378	-0.04728	0.89243	0.02922	0.05044	0.28902	0.06035	-0.2767	-0.06883
Spent Liquors	0.38784	-0.29759	-0.45207	0.03064	0.29582	0.01317	0.08922	-0.07004	0.45923	0.49424
Pyrolysis	0.3747	-0.15716	0.41891	0.16717	-0.18659	-0.01124	-0.56675	-0.47653	-0.07946	0.20474
Kinetics	0.3	-0.35684	0.12711	-0.26022	0.23346	0.68247	0.0166	0.23462	-0.31676	-0.15122
Biomass	0.4227	0.22361	-0.10658	0.10719	-0.47149	0.04465	-0.26811	0.60528	0.26269	-0.1301
Thermal Affects	0.1795	-0.08834	0.68784	-0.09519	-0.01284	-0.26503	0.48674	0.2714	0.13626	0.27972
Kraft Total	-0.11246	0.34285	0.24327	0.21369	0.70993	0.01633	-0.39751	0.279	0.14898	0.05146
Paper and Pulp Total	0.07609	0.51012	0.14927	0.08812	-0.08157	0.55632	0.30589	-0.37855	0.38475	-0.05384
Combined Cycle Power Plants with Gas Turbines	0.27255	0.5383	-0.18264	-0.13641	0.01958	-0.03989	0.08894	0.01848	-0.58966	0.47545

Table 15: Summary Statistics for the ten-variable system. Note that these values are used in the program in the appendix for finding the categorization of sub-keywords into each of the ten components.

	Obs	Mean	Std. Dev	Min	Max
Gassification	410	3.91E-09	1.325444	-1.31231	5.593557
Combustion	410	6.51E-09	1.279778	-3.2231	5.380592
Spent Liquors	410	2.50E-09	1.104598	-2.3642	4.960219
Pyrolysis	410	-2.01E-09	1.018413	-2.01284	2.662612
Kinetics	410	6.29E-09	0.966643	-3.03021	2.948821
Biomass	410	1.31E-10	0.901669	-2.15332	2.529387
Thermal Affects	410	6.03E-09	0.884977	-2.97328	2.966537
Kraft Total	410	-1.16E-09	0.844198	-3.03018	3.486636
Paper and Pulp Total	410	4.72E-10	0.795958	-3.22984	2.397334
Combined Cycle Power Plants with Gas Turbines	410	4.12E-10	0.686481	-2.01277	3.085398

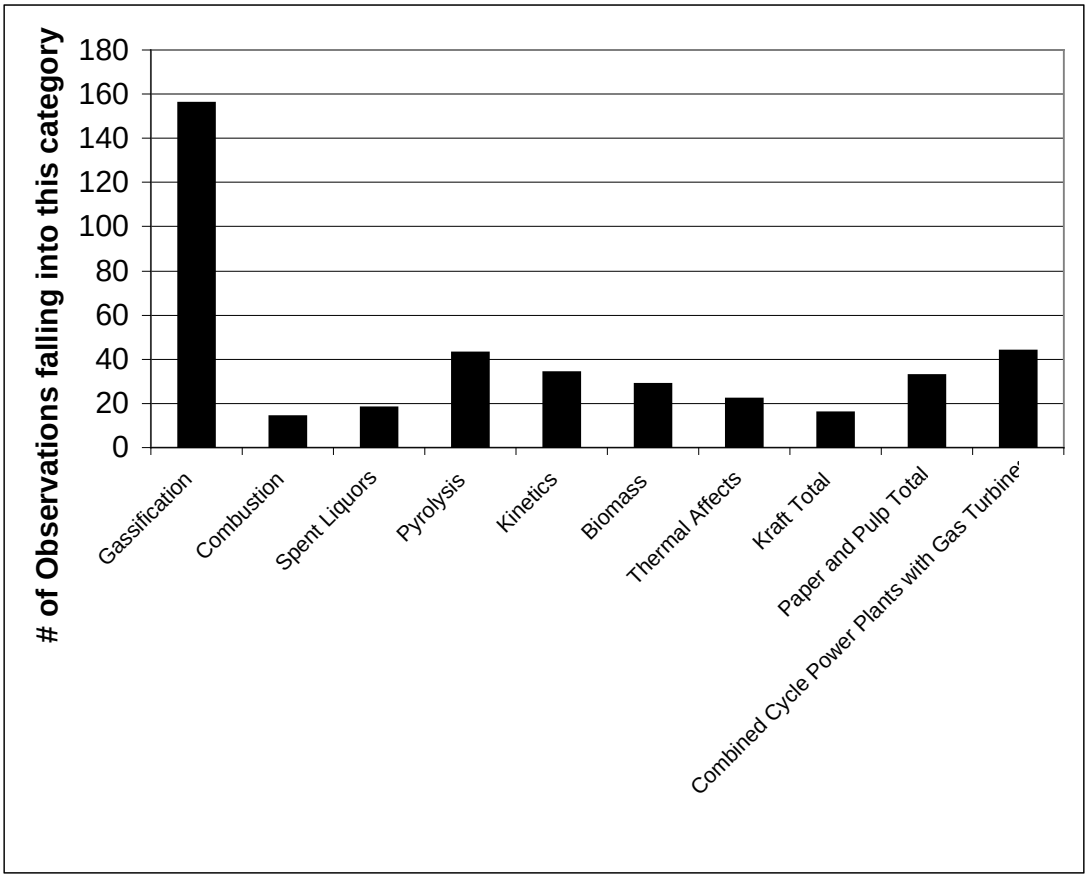

Figure 20: Histogram of number of matches between the ten-keyword variables and the sub-keywords.

Figure 20 indicates that the variables Combustion, Spent Liquors, and Kraft Total should be removed from the analysis for the second principal component iteration. The same method that was used on the first iteration of principal component analysis was also used on the second, except in this case there were seven categories of keywords with the same 400+ keyword list of sub-keywords. The code used to generate the principal components and their grouping is found in the appendix.

Table 16: Scoring for the seven-variable system

Scoring	Gasification	Pyrolysis	Kinetics	Biomass	Thermal effects	Paper and Pulp Total	Combined Cycle Power Plants with Gas Turbines
Gassification	0.46717	0.16654	0.00745	0.27889	-0.5983	-0.51582	-0.22836
Pyrolysis	0.20164	0.5446	0.16025	-0.50008	-0.38368	0.45159	0.18936
Kinetics	0.0199	0.52725	0.43504	0.6357	0.31512	0.16997	0.00622
Biomass	0.42707	0.00753	0.43386	-0.45607	0.50215	-0.40686	-0.06026
Thermal Affects	0.12037	0.50079	-0.72826	-0.02662	0.33334	-0.22932	0.19987
Paper and Pulp Total	0.52962	-0.15395	-0.25745	0.0676	0.17959	0.50985	-0.57684
Combined Cycle Power Plants with Gas Turbines	0.51315	-0.3509	-0.01115	0.23399	0.00177	0.15174	0.73187

Table 17: Eigenvectors for the seven-variable system

Eigenvectors	Gasification	Pyrolysis	Kinetics	Biomass	Thermal effects	Paper and Pulp Total	Combined Cycle Power Plants with Gas Turbines
Gassification	0.46717	0.16654	0.00745	0.27889	-0.5983	-0.51582	-0.22836
Pyrolysis	0.20164	0.5446	0.16025	-0.50008	-0.38368	0.45159	0.18936
Kinetics	0.0199	0.52725	0.43504	0.6357	0.31512	0.16997	0.00622
Biomass	0.42707	0.00753	0.43386	-0.45607	0.50215	-0.40686	-0.06026
Thermal Affects	0.12037	0.50079	-0.72826	-0.02662	0.33334	-0.22932	0.19987
Paper and Pulp Total	0.52962	-0.15395	-0.25745	0.0676	0.17959	0.50985	-0.57684
Combined Cycle Power Plants with Gas Turbines	0.51315	-0.3509	-0.01115	0.23399	0.00177	0.15174	0.73187

Table 18: Eigenvalue and difference statistics for the seven-variable system analysis

	Eigenvalue	Difference	Proportion	Cumulative
Gassification	1.90848	0.49959	0.2726	0.2726
Pyrolysis	1.4089	0.50542	0.2013	0.4739
Kinetics	0.90347	0.05679	0.1291	0.603
Biomass	0.84669	0.09852	0.121	0.7239
Thermal Affects	0.74817	0.05202	0.1069	0.8308
Paper and Pulp Total	0.69615	0.208	0.0994	0.9303
Combined Cycle Power Plants with Gas Turbines	0.48815	.	0.0697	1

Table 19: Summary Statistics for the seven-variable system. Note that these values are used in the program in the appendix for finding the categorization of sub-keywords into each of the ten components.

Variable	Observations	Mean	Std. Dev.	Min	Max
Gasification	382	1.03E-08	1.381478	-0.841738	6.353162
Pyrolysis	382	6.75E-09	1.186969	-2.328518	5.331581
Kinetics	382	1.17E-09	0.9505123	-3.984229	3.631588
Biomass	382	-3.95E-09	0.9201565	-3.584742	2.590187
Thermal Affects	382	2.83E-09	0.8649666	-2.293247	3.217453
Paper and Pulp Total	382	1.53E-09	0.8343546	-2.638615	2.595729
Combined Cycle Power Plants with Gas Turbines	382	-3.64E-10	0.6986747	-2.96395	2.694185

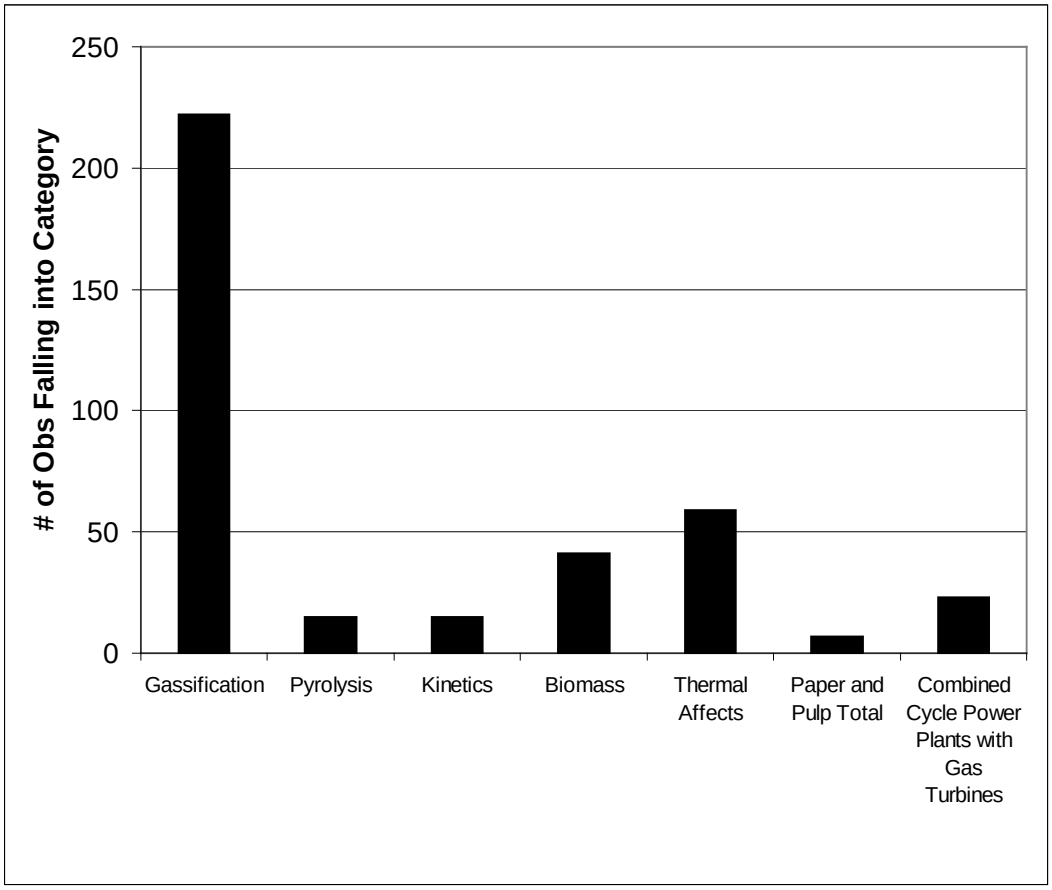

Figure 21: Histogram of number of matches between the seven-keyword variables and the sub-keywords.

Results

The result of this analysis is that the three areas that have the highest amount of observations fitting into the remaining categories are Gasification, Biomass, and Thermal Effects. These three final groups were again iterated through the system and their results were used in the process of determining sub groups from the data observation points. The final three groups were iterated through the principal components analysis one final time to categorize all of the sub-categories into the three remaining categories. After the categories were created, an analysis of the sub-keywords was done. Keywords that match in both the correlation and in the principal component analysis are listed in the appendix. It is noted that from the original list of 410 sub-keywords, there were only a total of 27 that were both a match in the VantagePoint program and in the principal component analysis.

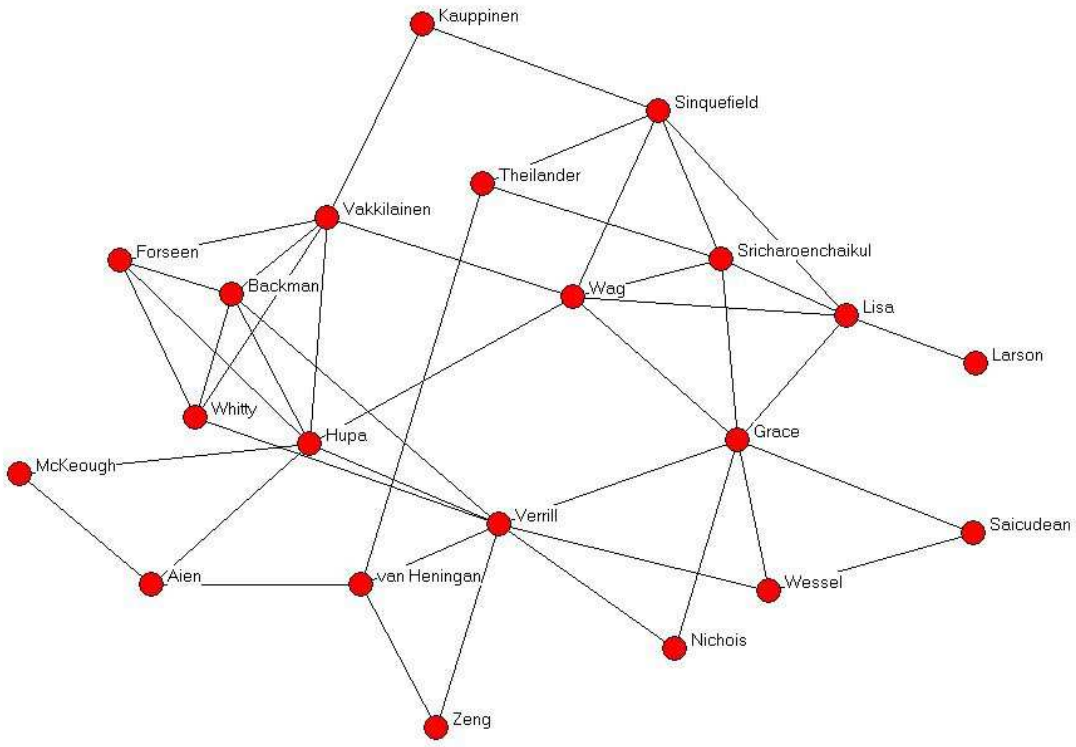

Figure 22: Total map of individuals working in the BLG arena and who also have published papers relating to Gasification, BioMass, or Thermal Effects.

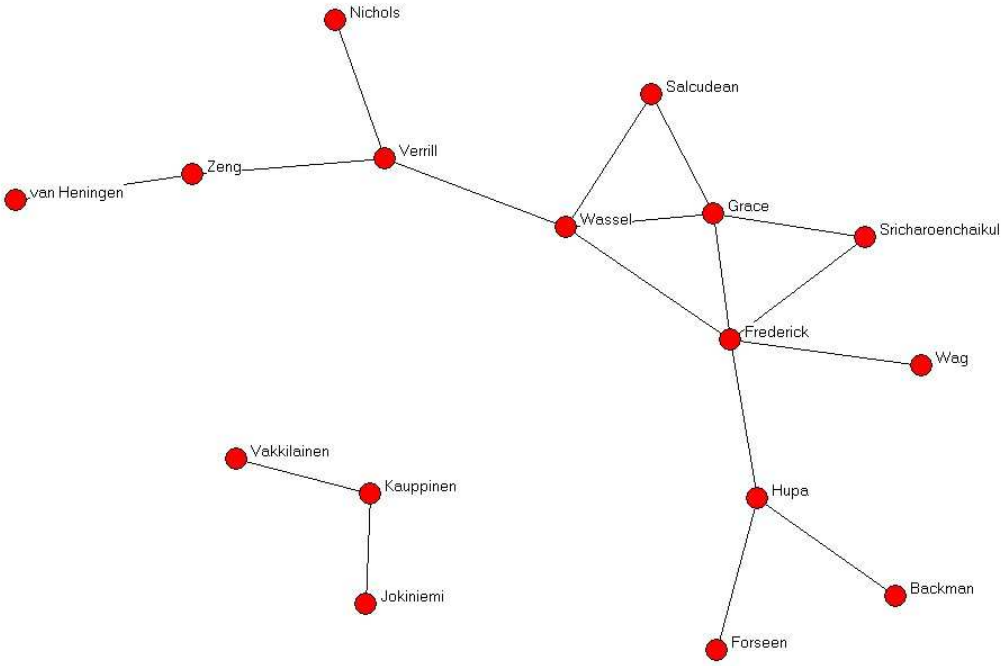

Figure 23: Map of Gasification authors.

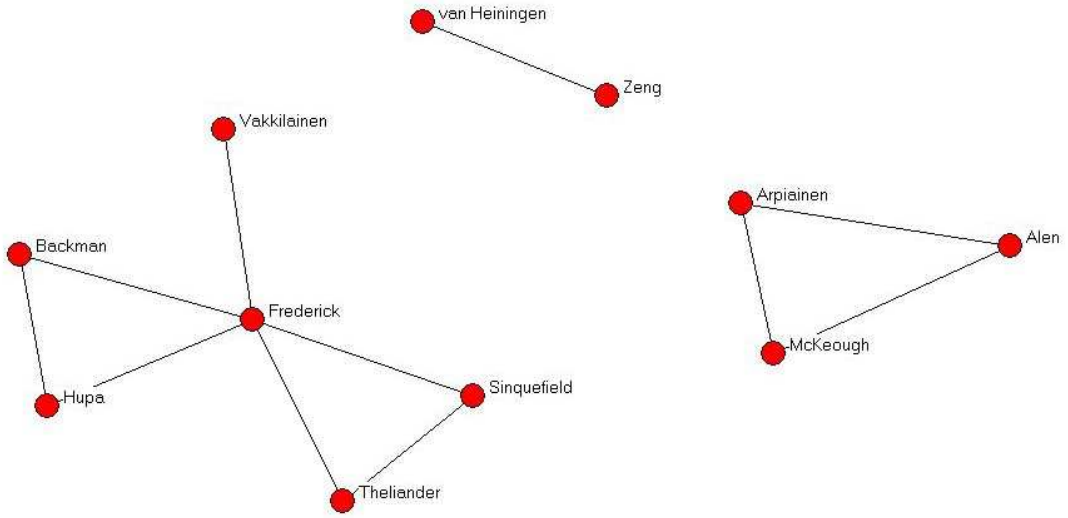

Figure 24: Map of Biomass authors.

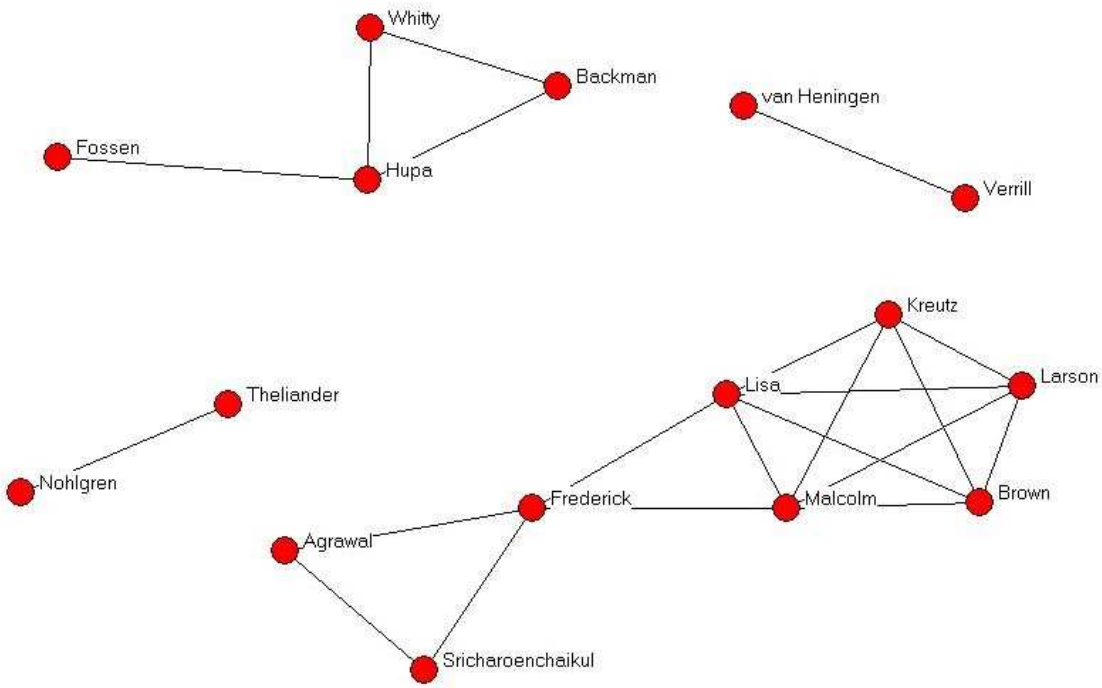

Figure 25: Map of Thermal Effects authors.

Now that maps of individuals and their co-authors are known and mapped from the three main principal components and their combination, the next step will be to find a means for finding probability of output based on previous output. From the principal

components analysis in the first part of this section, it is determinable which components sub-keywords fall into which of the three principal components. Also, programs such as Vantage Point give author statistics as to which authors published articles containing each of the keywords.

From this data, the number of sub-keywords for each author was summed for both the total list of all sub-keywords and for those sub-keywords that fall into each of the three principal component areas. Each author was compared to the most productive author in the total area or in the principal component area, and a percentage of their individual productivity to the most productive researcher was calculated. With each individual's productivity given as a percentage, all that is needed to calculate the percentage of output for a research collaboration would be to multiply the percentages of each researcher with every other researcher. Tables 20, 21, 22, and 23 contain the data for each of these probabilities of output.

Table 20: Co-authorship probabilities based on data for all of the three remaining principal components combined based on past output

	Hupa	Frederick	van Heiningen	Grace	Backman	Vakkilainen	McKeough	Forssen	Alen	Verrill
Hupa	1	0.948718	0.615385	0.461538	0.410256	0.358974	0.307692	0.307692	0.282051	0.282051
Frederick		0.948718	0.583826	0.43787	0.389218	0.340565	0.291913	0.291913	0.267587	0.267587
Van Heiningen			0.615385	0.284024	0.252465	0.220907	0.189349	0.189349	0.17357	0.17357
Grace				0.461538	0.189349	0.16568	0.142012	0.142012	0.130178	0.130178
Backman					0.410256	0.147272	0.126233	0.126233	0.115713	0.115713
Vakkilainen						0.358974	0.110454	0.110454	0.101249	0.101249
McKeough							0.307692	0.094675	0.086785	0.086785
Forssen								0.307692	0.086785	0.086785
Alen									0.282051	0.079553
Verrill										0.282051

Table 21: Co-authorship probabilities based on data for Gasification principal component category based on past output

	Hupa	Frederick	van Heiningen	Grace	Backman	Vakkilainen	McKeough	Forssen	Alen	Verrill
Hupa	1	0.764706	0.588235	0.382353	0.411765	0.294118	0.117647	0.323529	0.147059	0.205882
Frederick		0.764706	0.449827	0.292388	0.314879	0.224913	0.089965	0.247405	0.112457	0.157439
Van Heiningen			0.588235	0.224913	0.242215	0.17301	0.069204	0.190311	0.086505	0.121107
Grace				0.382353	0.157439	0.112457	0.044983	0.123702	0.056228	0.07872
Backman					0.411765	0.121107	0.048443	0.133218	0.060554	0.084775
Vakkilainen						0.294118	0.034602	0.095156	0.043253	0.060554
McKeough							0.117647	0.038062	0.017301	0.024221
Forssen								0.323529	0.047578	0.066609
Alen									0.147059	0.030277
Verrill										0.205882

Table 22: Co-authorship probabilities based on data for Biomass principal component category based on past output

	Hupa	Frederick	van Heiningen	Grace	Backman	Vakkilainen	McKeough	Forssen	Alen	Verrill
Hupa	0.25	0.27	0.25	0.75	0	1	0.25	0	0.25	0.25
Frederick		0.75	0.1875	0.5625	0	0.75	0.1875	0	0.1875	0.1875
Van Heiningen			0.25	0.1875	0	0.25	0.0625	0	0.0625	0.0625
Grace				0.75	0	0.75	0.1875	0	0.1875	0.1875
Backman					0	0	0	0	0	0
Vakkilainen						1	0.25	0	0.25	0.25
McKeough							0.25	0	0.0625	0.0625
Forssen								0	0	0
Alen									0.25	0.0625
Verrill										0.25

Table 23: Co-authorship probabilities based on data for Thermal Effects principal component category based on past output

	Hupa	Frederick	van Heiningen	Grace	Backman	Vakkilainen	McKeough	Forssen	Alen	Verrill
Hupa	0.5	0.5	0.375	0.25	0.25	0	0.875	0.125	0.625	0.375
Frederick		1	0.375	0.25	0.25	0	0.875	0.125	0.625	0.375
Van Heiningen			0.375	0.09375	0.09375	0	0.328125	0.046875	0.234375	0.140625
Grace				0.25	0.0625	0	0.21875	0.03125	0.15625	0.09375
Backman					0.25	0	0.21875	0.03125	0.15625	0.09375
Vakkilainen						0	0	0	0	0
McKeough							0.875	0.109375	0.546875	0.328125
Forssen								0.125	0.078125	0.046875
Alen									0.625	0.234375
Verrill										0.375

Discussion

For each of these alliances in each of the principal component areas of study and as a whole, this method seems to be strong for calculating future output in collaborative strategies between individuals. While the method used to calculate the percentages of connection between researchers was not elaborate, future studies can implement more advanced methods for finding more accurate probabilities of output based on a larger sample of inputs other than past performance in publishing in a certain area.

This crude bibiometric analysis is easily implemented for use in funding co-authorship alliances and in places where a specific research "task" needs to be catalyzed. For instance, if the government were interested in funding a research alliance between two research disciplines with two individuals in each of the groups, the network map which is created through the principal components analysis would be a good first pass at visualizing who is working with whom and who are the central characters in the research system, while the research output weightings help determine which authors are going to be potentially productive in the respective research environments.

Conclusion

There were four things that were accomplished in this paper. The first attainment was to introduce terms and equations in social network analysis to those people who may be unfamiliar to the topic. Most of these equations exist as metrics that are commonly used to find characteristics of social networks in the static case. This does not mean that these measures can only be used to the baseline of a graph of nodes though. In fact, at any point during a research simulation a snapshot of a graph can be taken and statistics can be run to get an instantaneous estimation of how a network is performing. This leads to the future possibility of social networks and their measures being used in computational policy where instantaneously courses of action are made and broken by adhering to certain performance standards.

The second task (performed on the Castillo dataset (2000 and 2003)) was to show how social networks analysis could be used to determine optimal connections (or injections) to be made and (hopefully) predict features in graphs based on these injections. Most of the basis for the policy injections is from economic or policy theory, and the basis for the connection probabilities in the datasets is grounded in economic theory of increasing or constant returns to scale.

While the dataset did perform as was expected, there were issues relating to the dataset. The first concern with the data is that there is no information concerning any of the nodes or connections. While each node in the system represents a real company that was funded through venture capital, there is no knowledge as to how the company performed with the funding or even the scale of the funding to the companies. The second

concern came from the model of simulation on the data. While the probabilities of connection were based on longstanding economic theory, because little knowledge is known about the connections and nodes except that there exists a connection between two nodes, this will not give enough information to prescribe a policy on the system as a whole.

The third task of this paper was to give an example of a social network of academic collaboration and show how policy could be used to help speed up research tasks. Here more information is known about each researcher and each connection between those researchers than in the Castillo dataset. Still, the connections that are injected upon these groups are based on the same logic as those in the Castillo dataset and rely more on the topography of the graph. Also, there is not enough information in this dataset to give a reliable means for predicting future outcome since all that is known is that two individuals have worked together in the past to co-author a paper.

The final task of this paper is to give a means for finding an individual's potential for future potential through a combination of bibliometric analysis, social network analysis, and regression analysis. The method used in this paper is not as elaborate as it could be if it were to include more variables such as institutional affiliation, number of years in researching the specific field, and keyword matching between potential unpaired authors. Still, this algorithm does give a means for reducing a set of research components and finding weights of connection for researchers in those components.

Most of the tools for analyzing social networks are becoming standardized in both language and method, but there is work to do in the area of dynamic social networks, especially in the topic of finding accurate weights for nodes and connections and accurate

models for the growth of networks. Concerning social networks and public policy, there is still work to be done to confirm that there is a one-to-one correspondence between existing policy theory and network modeling. This can be accomplished through more extensive data collection and model implementation in real world scenarios.

An asset of this approach to policy is that the data needed for finding these stronger weights already exists in some form in other parallel research, so there is not much additional labor that is required to arrive at these results. Also, it is quite simple to generate results from data using network analysis, so the lag between data collection and result analysis can be minimized. This new design for policy analysis will also give hope to those researchers who are in the business of data collection because it will show that the fruits of their labor can be extended into a new area of study.

There is an old adage from the business community that states “its not what you know but who you know”. Social network analysis will not be taking that strong of a stance, but it will aid in showing how position in a network can be a critical metric for finding ones potential. Although social network analysis alone should not be taken as a means for developing policy, there is hopefully enough motivation from this work to show how policy managers and policy analysts can implement it along with other traditional studies of output and econometrics to produce even stronger foundations for judgment.

Appendix

```
#Code #
#Code for generating the probability connections in MATLAB based on Connectivity
for i=1:length(A)
   for j=i:length(A)
      if i~=j
         if A(i,j)==1
            if nnz(A(:,i))>7
               A(i,j)=.6;
               A(j,i)=.6;
            end
            if nnz(A(:,i))>2&&nnz(A(:,i))<=7
               A(i,j)=.3;
               A(j,i)=.3;
            end
            if nnz(A(:,i))<=2
               A(i,j)=.1;
               A(j,i)=.1;
            end
         end
      end
   end
end

#Code #
#Code for generating the probability connections in MATLAB based on Random
(constant returns)
for i=1:104
   for j=i:104
      if i~=j
         if A(i,j)==1
            bob=rand;
            if bob<=0.33
               A(i,j)=.1;
               A(j,i)=.1;
            elseif bob>0.33&bob<=0.66
               A(i,j)=.3;
               A(j,i)=.3;
            else
               A(i,j)=.6;
               A(j,i)=.6;
            end
         end
      end
   end
```

```
end

#Code #
#Code for finding the L/C ratio of a system of nodes.  This code does contain
contingencies for determining an L/C ratio even if there is not total accessibility in the
system, but that feature is not used to report any values in this paper.

jump=ones(length(A));
for i=1:length(A)
   totjump(i)=0;
   leftout(i)=0;
   for j=1:length(A)
      ANEW=A;
      while jump(i,j)<=19
         if ANEW(i,j)==0
            ANEW=A*ANEW;
            jump(i,j)=jump(i,j)+1;
         else
            break
         end
      end
      if jump(i,j)==20
         leftout(i)=leftout(i)+1;
      else
         totjump(i)=totjump(i)+jump(i,j);
      end
   end
   reachablenodes(i)=length(A)-leftout(i);
   averagejumps(i)=totjump(i)/reachablenodes(i);
   averagejumpstrength(i)=averagejumps(i)*(reachablenodes(i)/length(A));
end

for i=1:104
averagerealjump(i)=(totjump(i))/(reachablenodes(i)/104);
end
ave=mean(averagerealjump)

#Code #
#This code was used for running the simulation on the 104 node system on the Castillo
dataset#

%clear all
runs_baby=100; %number of times to iterate through the original graph
left_out=zeros(1,500); %initialize the left out matrix

for runbaby=1:runs_baby
```

```
   clear twostep_length maxcov twostep_bb P num_row num_col %clear used variables
between runs
   iter=1; %number of times to iterate through the system
   prob_mat=[.1 .3 .6]; %probabilities of the system to use (not used in this program)
   P=A; %make a new matrix P to store all the information of A and then add to it
   num_row = size(P,1); %set up number of rows in matrix
   num_col = size(P,2); %set up number of columns in the matrix
   tests=3; %number of times to test for connectivity
   tally_total_num=0; %sets the number of tallys to how the system responded
   tally_total_dem=0; %sets the total number of tallys in the system
   tally_total_left_side_num=0;
   tally_total_right_side_num=0;
   tally_total_left_dem=0;
   tally_total_right_dem=0;
   tally0=0; %initialize tally=0
   tally1=0; %initialize tally=1
   tally2=0; %initialize tally=2
   tally3=0; %initialize tally=3


   for iteration=1:1:iter
      i=1;
      while i<=num_row %go through the upper half of the matrix
         j=1;
         while j<=num_col %still going through the upper half of the matrix
            if A(i,j)==.1|A(i,j)==.3|A(i,j)==.6 %want to find values in the upper half of the
matrix that are neither zero or one
               tally=0;% create a 1*3 matrix to store whether or not a connection is made
               for k=1:tests % three tests for probability
                  rr=rand;
                  if rr<=A(i,j)
                     tally=tally+1;
                     if A(i,j)==.1
                        P(i,j)=.3; %shift probbailities up if connection is made
                        P(j,i)=.3; %shift probbailities of the symmetric part of the matrix too
                     end
                     if A(i,j)==.3
                        P(i,j)=.6; %shift probbailities up if connection is made
                        P(j,i)=.6; %shift probbailities of the symmetric part of the matrix too
                     end
                     if A(i,j)==.6
                        P(i,j)=.6; %shift probbailities up if connection is made
                        P(j,i)=.6; %shift probbailities of the symmetric part of the matrix too
                     end
                  end
                  if rr>A(i,j)
```

```
            if A(i,j)==.1
               P(i,j)=.1; %shift probbailities down if no connection is made
               P(j,i)=.1; %shift probbailities of the symmetric part of the matrix too
            end
            if A(i,j)==.3
               P(i,j)=.1; %shift probbailities down if no connection is made
               P(j,i)=.1; %shift probbailities of the symmetric part of the matrix too
            end
            if A(i,j)==.6
               P(i,j)=.3; %shift probbailities down if no connection is made
               P(j,i)=.3; %shift probbailities of the symmetric part of the matrix too
            end
         end
      end
      if tally==0 %no connnection made any time
         P(i,j)=0; %remove connection
         P(j,i)=0; %remove conjugate connection
      elseif tally==tests
         colspace=size(P,1);
         P(i,colspace+1)=.3; %add the new connection to the matrix
         P(j,colspace+1)=.3; %add the new connection to the matrix
         P(colspace+1,j)=.3; %add the new connection to the matrix
         P(colspace+1,i)=.3; %add the new connection to the matrix
         P(colspace+1,colspace+1)=1; %node needs to connect to itself
      end
      tally_total_num=tally_total_num+tally;
      tally_total_dem=tally_total_dem+3;
      if i<40|i==98|i==99|i==100|i==101|i==102|i==103|i==104
         tally_total_left_side_num=tally_total_left_side_num+tally;
         tally_total_left_dem=tally_total_left_dem+3;
      end
      if i>40&i<97
         tally_total_right_side_num=tally_total_right_side_num+tally;
         tally_total_right_dem=tally_total_right_dem+3;
      end
      if tally==0
         tally0=tally0+1;
      elseif tally==1
         tally1=tally1+1;
      elseif tally==2
         tally2=tally2+1;
      else
         tally3=tally3+1;
      end
   end
   j=j+1;
```

```
        end
        i=i+1;
    end
    A=P; %copy P into matrix A
    num_row = size(P,1); %set up number of rows in matrix
    num_col = size(P,2); %set up number of columns in the matrix

end

nodes_left_side=0;
for y=1:42
    nodes_left_side=nodes_left_side+nnz(P(:,y));
end
nodes_left_side_average=nodes_left_side/40;

nodes_right_side=0;
for y=42:97
    nodes_right_side=nodes_right_side+nnz(P(:,y));
end
nodes_right_side_average=nodes_right_side/56;


TWOSTEP=P*P; %generate a matrix of values 2 steps away from each point
twostep_length = size(TWOSTEP,1); %get length of this matrix even though it should
be the same size as num_row or num_col
for twostepp=1:twostep_length
    twostep_bb(1,twostepp)=nnz(TWOSTEP(twostepp,:)); %generate a
1*twostep_length matrix that is the number of nonzero values in each row or column of
TWOSTEP
end

average_connection_two_step_away=nnz(TWOSTEP)/length(TWOSTEP);

nodes_twostep_left_side=0;
for y=1:42
    nodes_twostep_left_side=nodes_twostep_left_side+nnz(TWOSTEP(:,y));
end
nodes_twostep_left_side_average=nodes_twostep_left_side/40;

nodes_twostep_right_side=0;
for y=42:97
    nodes_twostep_right_side=nodes_twostep_right_side+nnz(TWOSTEP(:,y));
end
nodes_twostep_right_side_average=nodes_twostep_right_side/56;
```

```
   connections_run=(nnz(P)-length(P))/2;


%     TT=A;
%
%     for kkk=1:length(TT)
%        tt=0;
%        while nnz(A(:,kkk))~=length(A)&tt~=104;
%
%           TT=A*TT;
%           tt=tt+1;
%        end
%        jummps(kkk)=tt;
%     end


   total_left_out=0; %initialize the number of unattached nodes after going through an
iteration
   for jjj=1:size(A)
      if nnz(A(jjj,:))==1
         total_left_out=total_left_out+1; %this is the value of nodes that are removed from
the system
      end
   end


   maxcov=max(twostep_bb); %find the max number of coverage points in TWOSTEP
   per_covered=(maxcov-1)/twostep_length; %subtract one to get rid of diagonal
elements
   TOTALCOV=P^twostep_length; %take the matrix to the all points
   tot_coverage=(nnz(TOTALCOV)-twostep_length)/(twostep_length^2); %find
percentage of coverage minus points to themselves

   for lpp=1:length(A)
      if nnz(P(:,lpp))==1
         left_out(1,lpp)=left_out(1,lpp)+1;
      end
   end


   initial_nodes_left_out=0;
   for lppp=1:length(MAT)
      if nnz(A(:,lppp))==1
         initial_nodes_left_out=initial_nodes_left_out+1;
      end
   end
```

```
   ratio_conn_to_nodes=(nnz(P)-length(P))/(2*length(P));


%     average_jumps_total(1,runsbaby)=mean(jummps);%after simulation run, find the
average number of jumps to get from one spot to the next
%     min_jumps_total(1,runsbaby)=min(jummps);
%     place_min_jumps_total(1,runsbaby)=find(jumps==min_jumps_total);
%     max_jumps_total(1,runsbaby)=max(jummps);
%     place_max_jumps_total(1,runsbaby)=find(jumps==max_jumps_total);

   initial_nodes_left_out_tot(1,runbaby)=initial_nodes_left_out;
   tally0_tot(1,runbaby)=tally0; %get the number of zero connections in the graph
   tally1_tot(1,runbaby)=tally1; %get the number of one connections in the graph
   tally2_tot(1,runbaby)=tally2; %get the number of two connections in the graph
   tally3_tot(1,runbaby)=tally3; %get the number of three connections in the graph
   percentage_connections(1,runbaby)=tally_total_num/tally_total_dem; %find the
percentage of total connections made in the graph
   size_dist(1,runbaby)=twostep_length; %generate a matrix of values of number of
points for each run
   per_covered_dist(1,runbaby)=per_covered; %generate a matrix of values of size
distribution of percentage covered
   tot_coverage_dist(1,runbaby)=tot_coverage; %generate a matrix of values of total
coverage of the graphs for all points
   total_left_out_dist(1,runbaby)=total_left_out; %generate a matrix of values of number
of nodes left out after "iter" iterations
   total_output_dist(1,runbaby)=tally_total_num/tally_total_dem; %find the percentage
of total connections made in the graph
   left_side_output_total(1,runbaby)=tally_total_left_side_num/tally_total_left_dem;
   right_side_output_total(1,runbaby)=tally_total_right_side_num/tally_total_right_dem;
   total_new_nodes(1,runbaby)=length(A)-104;

average_connection_two_step_away_total(1,runbaby)=average_connection_two_step_a
way;
   nodes_left_side_total(1,runbaby)=nodes_left_side;
   nodes_left_side_average_total(1,runbaby)=nodes_left_side_average;
   nodes_right_side_total(1,runbaby)=nodes_right_side;
   nodes_right_side_average(1,runbaby)=nodes_right_side_average;
   nodes_twostep_left_side_total(1,runbaby)=nodes_twostep_left_side;
   nodes_twostep_left_side_average_total(1,runbaby)=nodes_twostep_left_side_average;
   nodes_twostep_right_side_total(1,runbaby)=nodes_twostep_right_side;

nodes_twostep_right_side_average_total(1,runbaby)=nodes_twostep_right_side_average;
   connections_total(1,runbaby)=connections_run;
   ratio_conn_to_nodes_total(1,runbaby)=ratio_conn_to_nodes;
```

```
   A=MAT; %redefine A just to make sure that it has not changed

end

subplot(3,1,1)
hist(per_covered_dist)
title('Distribution of Percentage covered 2 steps from the max coverage point')
subplot(3,1,2)
hist(size_dist)
title('Distribution of size of graphs')
subplot(3,1,3)
hist(tot_coverage_dist)
title('Distribution of total coverage of graphs')

mean_of_total_left_out=mean(total_left_out_dist)
std_of_total_left_out=std(total_left_out_dist)
mean_of_percentage_covered=mean(per_covered_dist)
std_of_percentage_covered=std(per_covered_dist)
mean_of_suze=mean(size_dist)
std_of_size=std(size_dist)
mean_of_total_coverage=mean(tot_coverage_dist)
std_of_total_coverage=std(tot_coverage_dist)

#Code#
#This code was used to generate the PCs in the ten component system#


factor var1-var10, pc
score f1 f2 f3 f4 f5 f6 f7 f8 f9 f10
summarize f1 f2 f3 f4 f5 f6 f7 f8 f9 f10


gen f1difpca = abs((f1-3.91e-09))/1.325444
gen f2difpca = abs((f2-6.51e-09))/1.279778
gen f3difpca = abs((f3-2.50e-09))/1.104598
gen f4difpca = abs((f4-(-2.01e-09)))/1.018413
gen f5difpca = abs((f5-6.29e-09))/0.9666426
gen f6difpca = abs((f6-1.31e-10))/0.9016685
gen f7difpca = abs((f7-6.03e-09))/0.884977
gen f8difpca = abs((f8-(-1.16e-09)))/0.8441978
gen f9difpca = abs((f9-4.72e-10))/0.7959581
gen f10difpca = abs((f10-4.12e-10))/0.6864809

gen typepca = 0
replace typepca = 1 if (f1difpca >= f2difpca & f1difpca >= f3difpca &
f1difpca >= f4difpca & f1difpca >= f5difpca & f1difpca >= f6difpca &
f1difpca >= f7difpca & f1difpca >= f8difpca & f1difpca >= f9difpca &
f1difpca >= f10difpca)
replace typepca = 2 if (f2difpca >= f1difpca & f2difpca >= f3difpca &
f2difpca >= f4difpca & f2difpca >= f5difpca & f2difpca >= f6difpca &
```

```
f2difpca >= f7difpca & f2difpca >= f8difpca & f2difpca >= f9difpca &
f2difpca >= f10difpca)
replace typepca = 3 if (f3difpca >= f1difpca & f3difpca >= f2difpca &
f3difpca >= f4difpca & f3difpca >= f5difpca & f3difpca >= f6difpca &
f3difpca >= f7difpca & f3difpca >= f8difpca & f3difpca >= f9difpca &
f3difpca >= f10difpca)
replace typepca = 4 if (f4difpca >= f1difpca & f4difpca >= f2difpca &
f4difpca >= f3difpca & f4difpca >= f5difpca & f4difpca >= f6difpca &
f4difpca >= f7difpca & f4difpca >= f8difpca & f4difpca >= f9difpca &
f4difpca >= f10difpca)
replace typepca = 5 if (f5difpca >= f1difpca & f5difpca >= f2difpca &
f5difpca >= f3difpca & f5difpca >= f4difpca & f5difpca >= f6difpca &
f5difpca >= f7difpca & f5difpca >= f8difpca & f5difpca >= f9difpca &
f5difpca >= f10difpca)
replace typepca = 6 if (f6difpca >= f1difpca & f6difpca >= f2difpca &
f6difpca >= f3difpca & f6difpca >= f4difpca & f6difpca >= f5difpca &
f6difpca >= f7difpca & f6difpca >= f8difpca & f6difpca >= f9difpca &
f6difpca >= f10difpca)
replace typepca = 7 if (f7difpca >= f1difpca & f7difpca >= f2difpca &
f7difpca >= f3difpca & f7difpca >= f4difpca & f7difpca >= f5difpca &
f7difpca >= f6difpca & f7difpca >= f8difpca & f7difpca >= f9difpca &
f7difpca >= f10difpca)
replace typepca = 8 if (f8difpca >= f1difpca & f8difpca >= f2difpca &
f8difpca >= f3difpca & f8difpca >= f4difpca & f8difpca >= f5difpca &
f8difpca >= f6difpca & f8difpca >= f7difpca & f8difpca >= f9difpca &
f8difpca >= f10difpca)
replace typepca = 9 if (f9difpca >= f1difpca & f9difpca >= f2difpca &
f9difpca >= f3difpca & f9difpca >= f4difpca & f9difpca >= f5difpca &
f9difpca >= f6difpca & f9difpca >= f7difpca & f9difpca >= f8difpca &
f9difpca >= f10difpca)
replace typepca = 10 if (f10difpca >= f1difpca & f10difpca >= f2difpca
& f10difpca >= f3difpca & f10difpca >= f4difpca & f10difpca >= f5difpca
& f10difpca >= f6difpca & f10difpca >= f7difpca & f10difpca >= f8difpca
& f10difpca >= f9difpca)

count if typepca==1
count if typepca==2
count if typepca==3
count if typepca==4
count if typepca==5
count if typepca==6
count if typepca==7
count if typepca==8
count if typepca==9
count if typepca==10
```

#Code #
#This code was used to generate the PCs in the seven component system#

```
factor var1 var4 var5 var6 var7 var9 var10, pc
score f1 f4 f5 f6 f7 f9 f10
summarize f1 f4 f5 f6 f7 f9 f10
```

```
gen f1difpca = abs((f1-(-8.20e-09)))/1.299407
gen f4difpca = abs((f4-(-8.18e-10)))/1.203501
gen f5difpca = abs((f5-(-5.50e-09)))/0.9663367
gen f6difpca = abs((f6-(-5.97e-09)))/0.9150532
gen f7difpca = abs((f7-6.01e-09))/0.9018631
gen f9difpca = abs((f9-3.64e-09))/0.8531149
gen f10difpca = abs((f10-(-2.30e-09)))/0.7421825

gen typepca = 0
replace typepca = 1 if (f1difpca >= f4difpca & f1difpca >= f5difpca &
f1difpca >= f6difpca & f1difpca >= f7difpca & f1difpca >= f9difpca &
f1difpca >= f10difpca)
replace typepca = 4 if (f4difpca >= f1difpca & f4difpca >= f5difpca &
f4difpca >= f6difpca & f4difpca >= f7difpca & f4difpca >= f9difpca &
f4difpca >= f10difpca)
replace typepca = 5 if (f5difpca >= f1difpca & f5difpca >= f4difpca &
f5difpca >= f6difpca & f5difpca >= f7difpca & f5difpca >= f9difpca &
f5difpca >= f10difpca)
replace typepca = 6 if (f6difpca >= f1difpca & f6difpca >= f4difpca &
f6difpca >= f5difpca & f6difpca >= f7difpca & f6difpca >= f9difpca &
f6difpca >= f10difpca)
replace typepca = 7 if (f7difpca >= f1difpca & f7difpca >= f4difpca &
f7difpca >= f5difpca & f7difpca >= f6difpca & f7difpca >= f9difpca &
f7difpca >= f10difpca)
replace typepca = 9 if (f9difpca >= f1difpca & f9difpca >= f4difpca &
f9difpca >= f5difpca & f9difpca >= f6difpca & f9difpca >= f7difpca &
f9difpca >= f10difpca)
replace typepca = 10 if (f10difpca >= f1difpca & f10difpca >= f4difpca
& f10difpca >= f5difpca & f10difpca >= f6difpca & f10difpca >= f7difpca
& f10difpca >= f9difpca)

count if typepca==1
count if typepca==4
count if typepca==5
count if typepca==6
count if typepca==7
count if typepca==9
count if typepca==10
```

#Code #
#This code was used to generate the PCs in the three component system#

```
factor var1 var2 var3, pc
score f1 f2 f3
summarize f1 f2 f3


gen f1difpca = abs((f1-(-1.03e-08)))/1.122863
gen f2difpca = abs((f2-4.62e-09))/0.9921773
gen f3difpca = abs((f3-(-3.56e-09)))/0.8687712

gen typepca = 0
replace typepca = 1 if (f1difpca >= f2difpca & f1difpca >= f3difpca)
replace typepca = 2 if (f2difpca >= f1difpca & f2difpca >= f3difpca)
replace typepca = 3 if (f3difpca >= f1difpca & f3difpca >= f2difpca)
```

```
count if typepca==1
count if typepca==2
count if typepca==3

#Categories#
#The following sub-keywords were found to match the three main
categories for those generated in Vantage Point and those matching
categories from the principal component analysis#
```

Gasification Sub-Categories
Pyrolysis and gasification behavior of black liquor under pressurized conditions
LIEKKI 1 - Vuosikirja 1997. Seurantaryhmaeraportit. LIEKKI 1 - tutkimusohjelman julkaisuluettelo 1993-1996. (LIEKKI 1 - Annual Review 1997. Reports of the review group. Publication list 1993-1996)
Industry's role in commercialization? The Agenda 1010 perspective
Commercializing black liquor and biomass gasifier/gas turbine technology
Growing power. Bioenergy technology from Finland
Possibilities for new black-liquor processes in the pulping industry: Energy and emissions
LIEKKI and JALO Combustion and fuel conversion. Evaluation of research programmes 1988-1990
Case study on simultaneous gasification of black liquor and biomass in a pulp mill
MTCI/Thermochem steam reforming process for solid fuels for combined cycle power generation
Proceedings of the Seminar on Power Production from Biomass
Optical pyrometric measurements of surface temperatures during black liquor char burning and gasification
Combined biomass and black liquor gasifier/gas turbine cogeneration at pulp and paper mills
Gasification of Biomass. Final Report Stage 6

BioMass Sub-Categories
Basic studies on black-liquor pyrolysis and char gasification
Influence of char formation conditions on pressurized black liquor gasification rates
Pyrolysis of black liquor in a pressurized free fall reactor and in a pressurized grid heater
Pyrolysis of black liquors from alkaline pulping of straw. Influence of a preoxidation stage on the char characteristics
Leaching of NaOH from 4:5-sodium-titanate produced in an autocausticization process: Kinetics and equilibrium
Hydrogen resources conversion of black liquor to hydrogen rich gaseous products
Carbon gasification of kraft black liquor solids in the presence of TiO₁ in a fluidized bed
Sulphur distribution during air gasification of kraft black liquor solids in a fluidized bed of TiO₁ particles
Rapid pyrolysis of kraft black liquor

Thermal Effects-Categories
Utilization of urban and pulping wastes to produce synthetic fuel via pyrolysis
Black liquor and biomass gasification combined cycle (Development and commercialization in USA)
Influence of ash deposit chemistry and structure on physical and transport properties
Biomass-gasifier/gas turbine cogeneration in the pulp and paper industry
Suovan ja ligniinin jalostaminen polttonesteiksi. Loppuraportti. (Conversion of potash soap and lignin into liquid fuels. Final report)
Black liquor gasification characteristics. 1. Formation and conversion of carbon-containing product gases
Research on black-liquor conversion at the Technical Research Centre of Finland

References

Aghion, P., Howitt, P., 1992. “A model of growth through creative destruction”. Econometrica 60, 323–351.)

Albert, J.H., and Chib, S. 1993. “Bayesian Analysis of binary and polychotomous response data,” *Journal of the American Statistical Association.* v88. *pp.* 669-679.

S. Branigan, H. Burch, B. Cheswick. “Mapping and Visualizing the Internet”, Usenix Security Symposium, 2000

Burt, R. S. (1992). Structural holes: The social structure of competition. Cambridge, MA: Harvard University Press.

Castilla, Emilio J., Hokyu Hwang. Mark Granovetter and Ellen Granovetter. 2000. "Social Networks in Silicon Valley." Pp. 218-247 (Chapter 11) in *The Silicon Valley Edge: A Habitat for Innovation and Entrepreneurship*, edited by Chong-Moon Lee, William F. Miller, Henry Rowen, and Marguerite Hancock. Stanford: Stanford University Press.

Castilla, Emilio J. 2003. “Networks of venture capital firms in Silicon Valley,” *International Journal of management Technology*. V25. #1-2. pp 113-135.

Chase, Ivan. 1980. “Social process and hierarchy formation in small groups: A comparative perspective,” *American Sociological Review*. v45.

deLeon, Peter. 1999. The Missing Link Revisited: Contemporary Implementation Research. *Policy Studies Review*. 17. 311-338.

deLeon, Peter and Linda deLeon. 2002. What Ever Happened to Policy Implementation? An Alternative Approach. *Journal of Public Administration Research and Theory*, 12, 4, 467-492.

Dietz, James. (2004). "Scientists and Engineers in Academic Research Centers-an Examination of Career Patterns and Productivity". PhD Dissertation at the Georgia Tech School of Public Policy

Dodds, P.S. and Watts, D. J. 2004 (forthcoming). “Universal Behavior in a Generalized Model of Contagion.” *Physical Review Letters.*

Dunne, J.A, R.J. Williams, and N.D. Martinez . 2002. “Network structure and biodiversity loss in food webs: robustness increases with connectance” . *Ecology Letters* 5:558-567. Also, Santa Fe Institute Working Paper 02-03-013

Ennis, James G. 1992. "The Social Organization of Sociological Knowledge: Modeling the Intersection of Specialties." *American Sociological Review. v*57. pp 259-265.

Farmer, Michael. 2004 “BLG and Social Networks” *Sloan Foundation* Paper

Faust, Katherine. 1997. "Centrality in Affiliation Networks" *Social Networks* v19. pp.157-191.

Frank, K. and Yasumoto, J. 1998. "Social Capital Within and Between Subgroups." *American Journal of Sociology.* v104. #3. pp. 642-86.

Fujita, M. 1993. “Monopolistic competition and urban systems,” *European Economic Review* v.37. p. 308-315.

J. Furman, Michael Porter and S. Stern (2002) “Determinants of National Innovation Capacity,” *Research Policy,* 31,6: 899-933.

Krugman, Paul. 1980. “Scale economies, product differentiation and the pattern of trade.” *American Economic Review*, v70. p. 950 - 959.

Heclo, Hugh. 1977. A Government of Strangers: Executive Politics in Washington (Brookings, 1977)

Heclo, Hugh. 1978. “Issue Networks and the Executive Establishment,” in *The New American Political System* ed. By A. King. Washington: American Enterprise System.

Hicks, D, A. Breitzman K. Hamilton, and F. Narin (2000) “Research Excellence and Patented Innovation,” *Science and Public Policy,* 27,5: 310-321.

Krugman, P. and Smith, A. 1994. *Empirical Studies of Strategic Trade Policy.* University of Chicago Press. Chicago, Ill.

Krugman, Paul, 1991. “Increasing Returns and Economic Geography”. *The Journal of Political Economy*, Vol. 99, No. 3. (Jun., 1991), pp. 483-499.

B. Lundvall, B. Johnson, E. Andersen, and B. Dalum (2002) “National Systems of Production, Innovation and Competence Building,” *Research Policy*, 31,2: 213-231.

Maynard-Moody, Stephen and Suzanne Leland, 1999. Stories from the Front-lines of Public Management: Street-level Workers as Responsible Actors. In H.G. Rainey, J.L Brudney, L.J. O’Toole, Jr. (eds) *Advancing Public Management: New Developments in Theory, Methods, and Practice.* Washington, DC: Georgetown University Press.

Moody, J. and White, DR. 2003. “Social Cohesion and Embeddedness: A Hierarchical Concept of Social Groups.” *American Sociological Review.* v68. #1. pp.1-25.

Murray, Fiona (2003) “Innovation as Co-Evolution of Scientific and Technological Networks: Exploring Tissue Engineering,” *Research Policy*, 31, 8-9: 1389-1403.

R. Nelson (2003) “On the Uneven Evolution of Human Know-how,” *Research Policy*, 32, 6: 909-922.

M. E. J. Newman. “The structure and function of complex networks”, *SIAM Review* 45, 167-256 (2003).

Rycroft, R. and D. Kash, (1992) "Technology Policy Requires Picking Winners," *Economic Development Quarterly*.

Romer, Paul. 1987. “Growth based on increasing returns due to specialization.” *American Economic Review*, v77. #2. p. 56-62.

Sabatier, Paul, Loomis, John, and Catherine McCarthy. “Hierarchical Controls, Professional Norms, Local Constituencies, and Budget Maximization: An Analysis of U.S. Forest Service Planning Decisions.” *American Journal of Political Science*, Vol. 39, No. 1. (Feb., 1995), pp. 204-242.

Salter and Ben Martin (2001) “The Economic Benefits of Public Funded Basic Research: A Critical Review,” Research Policy, 30, 3: 509-532.

Smith, T. & Stevens, G. 1999. "The Architecture of Small Networks: Strong Interaction and Dynamic Organization in Small Social Systems." *American Sociological Review*. v64: p. 403-420.

Strogatz, Steven. 1994. *Nonlinear Dynamics and Chaos. With Applications to\ Physics,Biology, Chemistry, and Engineering* (Addison-Wesley: Reading, Mass.)

Tijssen, R. (2002) “Science Dependence of Technologies: Evidence from Inventions and their Inventors,” *Research Policy*, 31,4: 509-526.

Wasserman, S. and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Watts, D.J. and Steven H. Strogatz. 1998. "Collective Dynamics of 'small-world' networks." *Nature*. v393. pp. 440-442.

Watts, D.J. 1999. *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton: Princeton University Press.

White, D. R. 2003. “Network Analysis, Social Dynamics and Feedback in Social Systems.” *Cybernetics and Systems*. v35. #2-3. pp. 173-192.

White, D.R. Powell, W.W. Owen-Smith, J. and Moody, J. 2003. “Network Models and Organization Theory: from embeddedness to ridge structure.” In *Computational and Mathematical Organization Theory*. eds. Alessandro Lomi and Phillipa Pattison.

Williamson, Oliver E. 1975. *Market and Hierarchies: Analysis and Antitrust Implications*. Free Press. New York, NY.

Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* Free Press. New York, NY.

Wolfram, S. 1983. "Statistical Mechanics of Cellular Automata," *Reviews of Modern Physics.* v55. pp. 601-644.